# KÉPESÍTŐFORDÍTÁS

# Legal statement

**Magyar Agrár- és Élettudományi Egyetem**

**Szent István Campus**

Vidékfejlesztés és Fenntartható Gazdaság Intézet

Idegen Nyelvi Tanszék

Szakfordító szakirányú továbbképzési szak

# TEXT ANALYSIS IN R

# SZÖVEGELEMZÉS R-BEN

Belső konzulens: Veresné Dr. Valentinyi Klára

tanszékvezető, egyetemi tanár

Külső konzulens: Dr. Temesi Ágoston

egyetemi docens

Agrár-és Élelmiszergazdasági Intézet

Készítette: Lakner Zoltán

Gödöllő

2024

*“Igyekeztem azon, hogy a deák bötünek értelmét hiven magyaráznám; a szóllásnak módgyát pedig úgy ejteném, hogy ne láttatnék deákból csigázott homályossággal repedezettnek, hanem oly kedvesen folyna, mint-ha először magyar embertül, magyarúl iratott vólna”*

Pázmány Péter: Kempis Tamásnak Chcistus követésérül négy könyvei

idézi: Hargittay (2016)

**Tartalomjegyzék**

**KÖSZÖNETNYILVÁNÍTÁS**

*Az angol nyevel elsajátításáért végzett, több mint négy évtizedes személyes küzdelmem legszebb és legtartalmasabb két évét tölthettem a MATE Agár-és Természettudományi Szakfordító képzésén. Nemcsak nyelvtudásban, de látóköröm bővítésében is nagyon sokat jelentettek ezek az újból iskolapadban töltött hónapok. Köszönöm csoporttásaim, dr Prokay Enikő és a MATE Tanulmányi Oszálya dolgozója, Vajdai Ánges határtalan támogatását. Tisztelettel, szeretettel és hálával gondolok a képzésben részt vevő valamennyi oktatóra: Heltai Pál professzor úrra, akinek életbölcsessége, a nyelvek és a oktatás iránt tisztelete és alázata, valamint elképesztő általános műveltsége számomra örök példakép marad, Nagy Mariannára és Urbán Miklósra akik egy világszínvonalú fordítástámogató eszköz elsajátítását könnyítették meg, Varga Erika docens asszonyra, aki lankadatlan türelemmel és megértéssel segített eligazodni az angol nyelv-és nyelvhelyesség útvesztőiben, és –mindenek előtt- **Veresné Valentinyi Klára** professzor asszonyra, akinek **szervezőkészsége, pedagógiai elkötelezettsége és végtelen segítőkészsége, embersége nélkül sem ez a nagyon hasznos, példaértékű képzés, sem a jelen dolgozat nem jöhetett volna létre**.*

**NYILATKOZAT**

a diplomadolgozat nyilvános hozzáféréséről és eredetiségéről

A hallgató neve: Lakner Zoltán

A Hallgató Neptun kódja: PBPUNX

A dolgozat címe: Text analysis in R

A megjelenés éve: 2024

A konzulens intézetének neve: VIDÉKFEJLESZTÉS ÉS FENNTARTHATÓ GAZDASÁG INTÉZET

A konzulens tanszékének a neve: IDEGEN NYELVI TANSZÉK

Kijelentem, hogy az általam benyújtott

diplomadolgozat egyéni, eredeti jellegű, saját szellemi alkotásom. Azon részeket, melyeket más szerzők munkájából vettem át, egyértelműen megjelöltem, és az irodalomjegyzékben szerepeltettem.

Ha a fenti nyilatkozattal valótlant állítottam, tudomásul veszem, hogy a záróvizsga-bizottság a záróvizsgából kizár és a záróvizsgát csak új dolgozat készítése után tehetek.

A leadott dolgozat, mely PDF dokumentum, szerkesztését nem, megtekintését és nyomtatását engedélyezem.

Tudomásul veszem, hogy az általam készített dolgozatra, mint szellemi alkotás felhasználására, hasznosítására a Magyar Agrár- és Élettudományi Egyetem mindenkori szellemitulajdon-kezelési szabályzatában megfogalmazottak érvényesek.

Tudomásul veszem, hogy dolgozatom elektronikus változata feltöltésre kerül a Magyar Agrár-és Élettudományi Egyetem könyvtári repozitori rendszerébe. Tudomásul veszem, hogy a megvédett és

- nem titkosított dolgozat a védést követően

- titkosításra engedélyezett dolgozat a benyújtásától számított 5 év eltelte után

nyilvánosan elérhető és kereshető lesz az Egyetem könyvári repozitori rendszerében.

Kelt: Gödöllő, 2024 év április hó 23. nap

_____

Hallgató aláírása

**NYILATKOZAT**

a szerzői engedélyről a szöveg oktatási célú felhasználásához

A lefordítanó szöveget 2022. december hó 10. napján töltöttem le a Taylor & Frnacis kiadó hivatalos weboldaláról. Az ezzel kapcsolatos szabályozás, mely elérhető a https://www.tandfonline.com/terms-and-conditions webhelyen (utolsó elérés: 2024.04.28) a következőket írja:

“All articles published by Taylor & Francis / Routledge under the Creative Commons Attribution License on an open-access basis are licensed by the respective authors of such articles for use and distribution by you subject to citation of the original source in accordance with the terms of the license under which the work was published (please check the license statement for the relevant article(s). No permission is required from the authors or publishers.

If you make any translations of Taylor & Francis and Routledge Open articles and Taylor & Francis and Routledge Open Select articles, for which a prior translation agreement with Taylor & Francis has not been established, you must prominently display the following statement on any such translation:

"This is an unofficial translation of a [Taylor & Francis and Routledge Open article / Taylor & Francis and Routledge Open Select article] that appeared in a Taylor & Francis publication. Taylor & Francis and / or the rightsholder has not endorsed this translation."

A jelen Képesítőfordítás első oldalán az előírásnak megfelelően feltüntettem ezt a kitételt.

**NYILATKOZAT**

a diplomadolgozat nyilvános hozzáféréséről és oktatási célú felhasználásá

A hallgató neve: Lakner Zoltán

A Hallgató Neptun kódja: PBPUNX

A dolgozat címe: Text analysis in R

A megjelenés éve: 2024

A konzulens intézetének neve: VIDÉKFEJLESZTÉS ÉS FENNTARTHATÓ GAZDASÁG INTÉZET

A konzulens tanszékének a neve: IDEGEN NYELVI TANSZÉK

Kijelentem, hogy a leadott dolgozat, mely PDF dokumentum, megtekintését, nyomtatását és oktatási célú használatát engedélyezem.

Kelt: Gödöllő, 2024 év április hó 23. nap

**FORDÍTÓI ELEMZÉS**

1. **A szöveg fontosabb adatai**

**Eredeti cím: Text analysis in R**

**Célnyelvi cím: Szövegelemzés R-ben**

**Szerző:** Welbers, Kasper; Van Atteveldt, Wouter; Benoit, Kenneth

**Az eredeti megjelenés helye és ideje:** Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. Communication Methods and Measures, 11 (4), 245-265. http://www.tandfonline.com/doi/10.1080/ 19312458.2017.1387238

**2 A szöveg témája, szakterülete:**

A szöveg interdiszciplináris megközelítésű, mert egyszerre foglalkozik nyelvészeti, adattudományi és számítástechnikai problémákkal.

**3. Kommunikációs szint**

***3.1 A fordítási utasítás***

A szöveghez konkrét fordítási utasítás nem tartozott.

***3.2 A szöveg megjelenési módja, kommunikációs színtere:*** írott formában on-line elérhető nyilvános szöveg

***3.6 Célközönség:*** a pedagógiai és szociológiai kutatás témakörben általánosan tájékozott, az értő olvasás képességével rendelkező, de a matematikai statisztikában és a számítástechnikai eszközök alkalmazásában kevésbé jártas szakember olvasó számára

***3.7 Szerkesztése elvárások, nehézségek:*** A szöveg igyekezett megkülönböztetni tipográfiailag a programneveket és a `parancsok nevét`, ez azonban nem volt következetes

**4. Műfaji jellemzés**

***4.1 A szöveg műfaja*:** tudományos publikáció szakemberek számára.

***4.2 A szöveg típusa**:* tartalomközpontú

***4.3 A szöveg retorikai célja, eszközei:*** magyarázás, kifejtés, leírás, információátadás, ismeretközlés.

***4.5 Írói attitűd, kifejezésmód**:* objektív

***4.6 A szöveg stílusa, nyelvezete, nehézségi szintje:*** erősen szaknyelvi, tudományos.

Például:

Eng: In order to be able to map all known characters to a single scheme, the Unicode standard was proposed, although it also requires a digital encoding format (such as the UTF-8 format, but also UTF-16 or UTF-32).

Hun: Az Unicode szabványt azért dolgozták ki, hogy képesek legyünk valamennyi ismert karaktert egyetlen rendszerbe foglalni, jóllehet ez is digitális kódformátum alkalmazását (ilyen például az UTF-8, az UTF-16 vagy az UTF-32) teszi szükségessé.

### *4.7 A szöveg szerkezete:*

A szöveg írói –célkitűzésük szerint- bevezető tanulmányt kívántak írni, ennél azonban lényegesen komplebb, számos részletet tartalmazó írásmű készült.

### *4.8 Grammatikai regiszter*

*Mondatok*

A szöveg viszonylag hosszú, kijelentő mondatokból áll, melyeket esetenként felbontottam.

*16. szegmens*

Eng.: Each package extends the functionality of the base R language and core packages, and in addition to functions and data must include documentation and examples, often in the form of vignettes demonstrating the use of the package.

Hu.: Minden egyes program tovább bővíti az R-nyelv és az alapprogramok felhasználási lehetőségeit. Az új funkciók és adatok hozzáadásához dokumentációra és példákra is szükség van, melyek gyakran címkék formánjában mutatják be a program alkalmazását.

*Szórendváltoztatás*

A mondat szórendjét gyakran átalakítottam, hogy megfeleljen a magyar nyelv szabályainak.

*22. szegmens.*

*Eng.:* There is a vast collection of dedicated text processing and text analysis packages, from low-level string operations (Gagolewski, 2017) to advanced text modeling techniques such as fitting Latent Dirichlet Allocation models (Blei, Ng, & Jordan, 2003; Roberts et al., 2014)—nearly 50 packages in total at our last count.

*Hu.:* A legutóbbi felmérésünk alapján összesen közel ötven programot magába foglaló, kifejezetten a szövegfeldolgozást és a szövegelemzést szolgáló hatalmas gyűjteménye van, az alacsony szintű karakterlánc-kezelési műveletektől (Gagolewski, 2017) az olyan magas szintű szöveg-modellezési eljárásokig, mint például a látens Dirichlet allokációs modellek (Blei, Ng, & Jordan, 2003; Roberts et al., 2014).

*Aktív és passzív szerkezetek*

Mivel a szöveg egy szoftver használatának lépéseit mutatja be, azaz az alkalmazás folyamatát, ezért a szövegben elsősorban aktív igéket találunk.

*33. szegmens:*

Eng.: The first step, *importing text*, covers the functions for reading texts from various types of file formats (e.g., txt, csv, pdf) into a *raw text corpus* in R.

Hu.: Az első lépés a *szöveg importálás*a, mely a szöveg különböző file-formátumokból (pl. txt, csv, pdf) a *nyers szöveg-testbe (korpuszba)* történő beolvasásához szükséges feladatokat foglalja magába az R-be.

*Határozói igenevek*

A magyar számítástechnikai szaknyelv gyakran használ határozói igeneveket, ezért használtm az igésítést (grammatikai átváltási művelet) is, de időnként használtam a határozói igeneveket.

*4. szegmens*

Eng.: In this teacher's corner, we address these barriers by providing an overview of general steps and operations in a computational text analysis project, and demonstrate how each step can be performed using the R statistical software

Hu.: Ebben a módszertani segédletben ezeket az akadályokat tárgyaljuk és átfogó áttekintést adunk egy számítógépes szövegelemzési projekt általános kérdéseiről és műveleteiről, bemutatva, hogy minden egyes lépés hogyan hajtható végre az R statisztikai szoftver felhasználásával.

*Módbeli segédigék*

A szöveg objektív, leíró jellege miatt, ritka a módbeli segédige

*9.szegmens*

Eng.: However, for researchers that are not well-versed in programming, learning how to use R can be a challenge, and performing text analysis in particular can seem daunting.

Hu.: A programozásban kevésbé járatos kutatók számára azonban kihívást jelenthet az R használata, és a szövegelemzés megvalósítása különösen ijesztőnek tűnhet

However, for researchers that are not well-versed in programming, learning how to use R can be a challenge, and performing text analysis in particular can seem daunting.

A programozásban kevésbé járatos kutatók számára azonban kihívást jelenthet az R használata, és a szövegelemzés megvalósítása különösen ijesztőnek tűnhet.

***4.9 Lexikai regiszter:***

idegenes alakok és tükörfordítás használata gyakori

*5. szegmens*

Eng.: As a popular open-source platform, R has an extensive user community that develops and maintains a wide range of text analysis packages.

Hu.: Népszerű, nyílt forrású platformként az R kiterjedt felhasználói közösséggel rendelkezik, mely a szövegelemző programprogramok széles spektrumát fejleszti és tartja fenn.

*Eng.:* Working with these different packages and their different interfaces and output can be challenging, especially if different file formats are used together in the same project.

*Hu.:* A munkavégzés ezekkel a különböző programokkal és azok egymástól eltérő interfészeivel, illetve kimeneteivel nehézséget jelenthet, különösen akkor, ha ugyanazon projekt keretében más-más fájlformátumokat alkalmazunk.

*Tükörfordítás*

*8.szegmens*

Eng.: Currently, one of the most popular environments for computational methods and the emerging field of "data science" is the R statistical software (R Core Team, 2017).

Hu.: Jelenleg a számítási módszerek és a feltörekvő „adattudomány" egyik legnépszerűbb környezete az R statisztikai sz oftver (R Core Team, 2017).

A mű magyar nyelvű fordítása számos kihívást jelentett. Ennek három fő oka volt:

- maga az R programozási környezet is rendkívül dinamikusan, a szó szoros értelmében óráról órára fejlődik.
- a tanulmány készítői nehéz feladatra vállalkoztak, amikor egy ennyire komplex problémakörről kíséreltek meg áttekintést adni;
- a mű szerzői számára sem volt mindig egyértelmű, milyen előzetes ismereteket tételezke fel az ovasótól. Ebből az is következik, hogy sok esetben szerepeltek a szövegben olyan utalások, melyek jelentős számítástechnikia, és matematikai statisztikai háttérismeretet tesznek szükségessé.

Ebből adódóan a szöveg számos kérdést vet fel a fordító szemszögéből. A teljesség igénye nélkül ezek közül ötöt emelek ki:

1. A „function" szó fordításának kérdése. Az R környezetben az egyes algoritmusok megvalósítása függvényekkel történik, ezért semmi hibát nem követnénk el ha függvényt írnánk a fordításban. A fordítás felhasználója viszont koránt sem biztos hogy ilyen mértékben ismeri az R és az R alapú programok belső világát, tekintve hogy a tanulmány elsősorban társadalomkutatók részére készült. Tőlük ez nem is várható, ebből adódóan megítélésem szerint egyértelműen zavaró lett volna hogyha mechanikusan a függvény szót

alkalmazom, mert a témában kevésbé jártas, szakemberek számára ez így nehezebben lett volna követhető. Ezért gondoltam jobbnak a funkció szó alkalmazását, melyik pontosabban fejezi ki azt hogy itt egy olyan eszközről beszélünk amelyik valamely feladat ellátására képes. Ennyi háttérismeret bőségesen elegendő a fordításban leírt kutatások elvégzéséhez. Tovább bonyolítja a dolgot, hogyha a „függvény" fordításnál maradtam volna, akkor ehhez jól illeszkedik a szövegben gyakran előfordulónk „argumentum" kifejezés mely lényegében a függvény független változóit, azaz a felhasználó által megadandó értékeket tartalmazza. Ha azonban ezt a szót beemeltem volna a fordításba, félő lett volna, hogyha matematikában és számítástechnikában nem feltétlenül jártas olvasók nem értették volna, milyen független változókról beszélek, ezért jobbnak láttam „megadandó értékek"/"megadandó paraméterek" szókapcsolatokkal le egyszerűsíteni a helyzetet.

2. A nagy dilemmát jelentett a „package" szó fordítása is. A magyar nyelvű szakirodalom egy része a „package" szót alkalmazza, mások a csomag szót alkalmazzák. Megítélésem szerintem egyik sem viszi közelebb a felhasználót a megértéshez. Itt ugyanis programokról beszélünk melyek letölthetőek és letöltés után az R-ben alkalmazhatóak. Elvben lehetséges lett volna a „programcsomag" szó használata is ez azonban már foglalt, hiszen a számítástechnikában például a matematikai statisztikában ezt olyan nagy integrált programokra használják megyek a legkülönbözőbb szolgáltatásokat nyújtanak a felhasználók részére például a témakörben maradva a fel magyar felhasználó azt hallja hogy programcsomagolás akkor tapasztalataim szerinte körülbelül egy akkora rendszerre gondol mint a SPSS vagy a SAS etünkben azonban nem erről van szó: A programjaiak méretei ennél általában lényegesen kisebb és sokkal kevésbé complex ezért véleményem szerint félrevezető lett volna ez a fordítás

3. A számítástechnikai szleg magyarra történő átültetése. Mit jelentett számomra az is hogy a szerzők pontosan milyen domekratúrát nevezéktant alkalmaznak nehézséget okozott például a globe fordítása megyek ma már elavultnak tekinthető a számításainkában és amennyire a csonkolt alkalmazott (Válas, 1985; Cserkó és Pásztor, 2023)

4. A rövidítések használata. A szerzők nem fordítanak figyelmet arra, hogy következetesen feloldjanak minden, a szövegben található rövidítést, ezért – a szöveghűség megtartása érdekében- külön rövidítésjegyzéket mellékeltem a fordításhoz.

5. Az angol nyelvű szövegek magyarra történő fordításának terminológiája napjainkban még nem egységes (Demeczky, 2008). Az esetleges félreértések elkerülése érdekében igyekeztem következetesen alkalmazni fordítási megoldásait

### *4.10 Átváltási műveletek*

#### *Grammatikai átváltási műveletek*

*A grammatikai átváltási műveletek széles skáláját alkalmaztam a fordításánál.*

*nyelvtani csere*

*13.szegmens*

Eng.: In contrast to most programming languages, R was specifically designed for statistical analysis, which makes it highly suitable for data science applications.

Hu.: Más programnyelvekkel ellentétben az R-t kifejezetten statisztikai elemzésre tervezték, ezért kiválóan alkalmas adattudományi alkalmazásokra.

*41.szegmens*

Eng.: The *analysis* section discusses four text analysis methods that have become popular in communication research (Boumans & Trilling, 2016) and that can be performed with a DTM as input.

Hu.: Az *elemzés* fejezetben a kommunikációkutatásban népszerű négy szövegelemzési módszert tárgyalunk ((Boumans & Trilling, 2016), melyeket a DTM-input felhasználásával hajthatunk végre.

***Lexikai átváltási műveletek***

*A lexikai átváltási műveleteket is gyakran használtam a fordításánál.*

*lexikai kihagyás*

*9.szegmens*

Eng.: However, for researchers that are not well-versed in programming, learning how to use R can be a challenge, and performing text analysis in particular can seem daunting.

Hu.: A programozásban kevésbé járatos kutatók számára azonban kihívást jelenthet az R használata, és a szövegelemzés megvalósítása különösen ijesztőnek tűnhet.

*lexikai betoldás*

***12. szegmens***

Eng: R is a free, open-source, cross-platform programming environment.

Hu: Az R szabadon felhasználható, nyílt forráskódú, platformfüggetlen programozási környezet.

*lexikai csere*

***4. szegmens***

Eng: In this teacher's corner, we address these barriers by providing an overview of general steps and operations in a computational text analysis project, and demonstrate how each step can be performed using the R statistical software.

Hu: Ebben a módszertani segédletben ezeket az akadályokat tárgyaljuk és átfogó áttekintést adunk egy számítógépes szövegelemzési projekt általános kérdéseiről és műveleteiről, bemutatva, hogy minden egyes lépés hogyan hajtható végre az R statisztikai szoftver felhasználásával.

***12. szegmens***

Eng: R is a free, open-source, cross-platform programming environment.

Hu: Az R szabadon felhasználható, nyílt forráskódú, platformfüggetlen programozási környezet.

***15. szegmens***

Eng.: One of the keys to R's explosive growth (Fox & Leanage, 2016; TIOBE, 2017) has been its densely populated collection of extension software libraries, known in R terminology as *packages*, supplied and maintained by R's extensive user community.

Hu.: Az R robbanásszerű bővülésének (Fox & Leanage, 2016; TIOBE, 2017), egyik kulcsa a nagyszámú kiegészítő szoftverkönyvtár, az úgynevezett *programok*, melyeket az R kiterjedt felhasználói közössége készít és tart fenn.

***Terminológia*:**

*Idegenszavak és idegenes alakok, terminusok gyakori átvétele*

A szöveg az R-szoftvert mutatja be, amely egy új, rendkívül gyorsan, „cunamiként" terjedő statisztikai program. Újszerűségének köszönhetően a statisztikai szoftver angol terminusait használja a számítástechnika, ezt nevezzük „tsunami-hatásnak".

*Például:*

Eng.: Not only can multiple files be references using simple, "glob"-style pattern matches, such as /myfiles/*.txt, but also the same command will recurse through sub-directories to locate these files.

Hu.: Nem csak több fájl hivatkozható egyegyszerű, joker karaktert tartalmazó „glob" keresőkifejezéssel, mint például a /myfiles/*.txt hivatkozás, hanem ugyanez a parancs ismételhető az alkönyvtárakban is az ilyen fájlok helyének meghatározására.

*Betűszavak*

A szövegben található angol betűszavakat a magyar szövegben is megtartottam, hiszen a számítástechnikában az angol betűszavak használatosak.

*37. szegmens*

Eng.: Finally, it is a common step to *filter and weight* the terms in the DTM.

Hu.: Végül gyakori az egyes kifejezések *szűrése és súlyozása* is a DTM-ben (document-term matrix).

*Helyesírás*

*Kisbetűs, nagybetűs írásmód*

A szövegfájlok elnevezéseit a magyar számítástechnikai szaknyelvben többnyire kisbetűs alakban (köznyelvi szó betűszóként) használjuk, de a fordított célnyelvi szövegben többnyire megmaradt a forrásnyelvi szövegben használt nagybetűs alak.

Eng.: R natively supports reading regular flat text files such as CSV and TXT, but additional packages are required for processing formatted text files such as JSON (Ooms, 2014), HTML, and XML (Lang & the CRAN Team, 2017), and for reading complex file formats such as Word (Ooms, 2017a), Excel (Wickham & Bryan, 2017) and PDF (Ooms, 2017b).

Az R alapértelmezetten támogatja az olyan formázatlan szövegfájlok beolvasását, mint például a CSV és a TXT, de további programok letöltése szükséges az olyan formázott szövegfájlok feldolgozásához, mint például a JSON (Ooms, 2014), a HTML vagy az XML és az olyan komplex fileformátumokhoz, mint pl. a Word (Ooms, 2017a), az Excel (Wickham & Bryan, 2017) és a PDF (Ooms, 2017b).

*Kivéve:*

*33. szegmens:*

Eng.: The first step, *importing text*, covers the functions for reading texts from various types of file formats (e.g., txt, csv, pdf) into a *raw text corpus* in R.

Hu.: Az első lépés a *szöveg importálás*a, mely a szöveg különböző file-formátumokból (pl. txt, csv, pdf) a *nyers szöveg-testbe (korpuszba)* történő beolvasásához szükséges feladatokat foglalja magába az R-be.

*Egybeírás, különírás*

A magyar terminológiában a dokumentum-kifejezés mátrix szószerkezetes szókapcsolat különírva terjedt el. Ezt általában így alkalmazzák. Más szerzők (pl. Boda &.Sebők, 2018) egybeírást alkalmaznak: dokumentum-kifejezésmátrix (document-term matrix – DTM), sőt Ring és Kiss (2020) az egész fogalmat egybeírják: „dokumentumkifejezés-mátrix". A probléma bonyolultságát tovább növeli, hogy a szövegelemzés tudománya ugyanezt a rövidítést (DTM) alkalmazza a Dinamic Topic Model megjelölésére is (Strelicz, 2022).

Én Németh (2022) írásmódját fogadom el.

## 5. A gépi fordítás és utószerkesztés tapasztalatai

A gépi fordítást a DeepL online fordító alkalmazással készítettem. A DeepL legtöbbször tükörfordítást alkalmazott.

### A gépi fordítás tanulságai

1. A gép sokszor tükörfordítítást alkalmaz, ami hibás fordítást eredményez. Részleges utószerkesztés szintjén többnyire elfogadható a gépi fordítás, de amennyiben a human fordítással egyenértékű fordítást szeretnénk kapni,akkor teljes utószerkesztést kell végeznünk. Ezekben az esetekben többnyire a mondatszórendet is át kell alakítani.

   Például:

Eng.: Furthermore, with the use of special matrix formats for sparse matrices, text data in a DTM format is very memory efficient and can be analyzed with highly optimized operations.

DeepL: Továbbá a ritka mátrixok speciális mátrixformátumainak használatával a DTM formátumú szöveges adatok nagyon memóriahatékonyak, és rendkívül optimalizált műveletekkel elemezhetők

Teljes utószerkesztés: A ritka mátrixok speciális mátrixformátumainak használatával a DTM -ben tárolt szöveges adatok nagyon hatékony memóriakezelést és optimalizált műveletvégzést tesznek lehetővé.

2. A gép nem ismeri a speciális szakszavakat, ezekt tükörfordítással fordítja, ezért ezeket részleges és teljes utószerkesztésnél és javítani szükséges, azaz lexikai cserét kell végezni.
   Például:

Eng.: Two of the most established text analysis packages in R that provide dedicated DTM classes are tm and quanteda.

DeepL: Az R két legelterjedtebb szövegelemző csomagja, amelyek dedikált DTM osztályokat biztosítanak, a tm és a quanteda

Teljes utószerkesztés: Az tm és a quanteda az R két legelterjedtebb szövegelemző programja, amelyekkel speciális DTM fájlok hozthatók létre

Eng.: The performance and flexibility of quanteda's dfm format lends us to recommend it over the tm equivalent.

DeepL: A quanteda dfm formátumának teljesítménye és rugalmassága miatt ajánljuk a tm megfelelőjével szemben

Teljes utószerkesztés: A quanteda dfm fájlját teljesítménye és rugalmassága miatt ennek használatát ajánljuk a tm megfelelő funkciójával szemben

3. A gép nem ismeri a szakmai kollokációkat.

Például:

Eng.: As such, tidytext does not strictly use a document term matrix, but instead represents the same data in a long format, where each (non-zero) value of the DTM is a row with the columns document, term, and count (note that this is essentially a triplet format for sparse matrices, with the columns specifying the row, column and value).

DeepL: Mint ilyen, a tidytext nem szigorúan dokumentum-terminus mátrixot használ, hanem ugyanezeket az adatokat egy hosszú formátumban ábrázolja, ahol a DTM minden (nem nulla) értéke egy sor, amelynek oszlopai a dokumentum, a terminus és a szám (megjegyzendő, hogy ez lényegében a ritka mátrixok triplett formátuma, ahol az oszlopok megadják a sort, az oszlopot és az értéket).

Teljes utószerkesztés: Mint ilyen, a tidytext nem szigorúan dokumentum-terminus mátrixot használ, hanem ugyanezeket az adatokat egy olyan hosszú formátumban tárolja, ahol a DTM minden (nem nulla) értéke egy sor, és melynek oszlopai a dokumentum, a terminus és a gyakoriság (megjegyzendő, hogy ez lényegében a ritka mátrixok triplett formátuma, ahol az oszlopokba renezett információk adják meg a sort, az oszlopot és az értéket).

4. A gép nem használ átváltási műveleteket, például a hosszú mondatokat nem bontja fel. Ezt teljes utószerkesztésénél érdemes elvégezni.

    Például:

Eng.: Of the two, the venerable tm package is the more commonly used, with a user base of almost 10 years (Meyer, Hornik, & Feinerer, 2008) and several other R packages using its DTM classes (DocumentTermMatrix and TermDocumentMatrix) as inputs for their analytic functions.

DeepL: A kettő közül a tiszteletre méltó tm csomag a leggyakrabban használt, közel 10 éves felhasználói bázissal (Meyer, Hornik, & Feinerer, 2008) és számos más R csomag használja a DTM osztályait (DocumentTermMatrix és TermDocumentMatrix) elemző függvényeik bemeneteként

Teljes utószerkesztés: A kettő közül a jóöreg tm program a leggyakrabban alkalmazott, közel 10 éves felhasználói tapasztalattal (Meyer, Hornik, & Feinerer, 2008). Az általa létrehozott DTM fájlokat (DocumentTermMatrix és TermDocumentMatrix) számos más R program is használja az elemző funkcióik bemeneteként

5. A gép nem használja  a betoldás, illetve explicitáció átváltási műveleteket. A kutatások szerint a human fordítás mindig hosszabb, mint a forrásnyelvi szöveg, mert a jobb érthetőség miatt a fordító gyakran explicitál, ezért az explicitációt a teljes utószerkesztésnél érdemes elvégezni.

Például:

Eng.: Its sparse DTM class, known as a dfm or document-feature matrix, is based on the powerful Matrix package (Bates & Maechler, 2015) as a backend, but includes functions to convert to nearly every other sparse document-term matrix used in other R packages (including the tm formats).

DeepL: Gyér DTM-osztálya, amelyet dfm vagy dokumentum-tulajdonságmátrix néven ismerünk, háttértárként a nagy teljesítményű Matrix csomagra (Bates & Maechler, 2015) épül, de tartalmaz olyan függvényeket, amelyekkel szinte minden más, más R-csomagokban használt gyér dokumentum-term mátrixra (beleértve a tm formátumokat is) konvertálható

Teljes utószerkesztés: Az programmal létrehozott, a ritka mátrixok kezelését lehetővé tevő DTM fájl, amelyet dfm vagy dokumentum-tulajdonságmátrix néven ismerünk, háttértárként a nagy teljesítményű Matrix programra (Bates & Maechler, 2015) épül, de tartalmaz olyan funkciókat is, amelyekkel szinte minden más, R-programban használt ritka dokumentum-term mátrix (beleértve a tm formátumokat is) is konvertálható

6. Időnként a gép meglepően jól,szinte hibátlanul fordít.

Például:

Eng.: This is a text analysis package that is part of the Tidyverse7—a collection of R packages with a common philosophy and format.

DeepL: Ez egy szövegelemző csomag, amely része a Tidyverse7 - egy közös filozófiájú és formátumú R-csomagok gyűjteményének.

Teljes utószerkesztés: Ez egy olyan szövegelemző program, amely része a Tidyversenek, azaz egy közös elven és formátumon alapuló R-programok gyűjteményének.

7. Megállapítom, hogy a részleges utószerkesztésnél elsősorban terminusokat változtattam.

Eng.: For an overview of text analysis approaches we build on the classification proposed by Boumans and Trilling (2016) in which three approaches are distinguished: counting and dictionary methods, supervised machine learning, and unsupervised machine learning.

DeepL: A szövegelemzési megközelítések áttekintéséhez a Boumans és Trilling (2016) által javasolt osztályozásra építünk, amelyben három megközelítést különböztetünk meg: számláló és szótári módszerek, felügyelt gépi tanulás és felügyelet nélküli gépi tanulás.

Teljes/részleges utószerkesztés: A szövegelemzési megközelítések áttekintéséhez a Boumans és Trilling (2016) által javasolt osztályozásra építünk, amelyben három módszert különböztetünk meg: számláló és szótári módszerek, felügyelt gépi tanulás és felügyelet nélküli gépi tanulás.

8. Stiláris javítást csak a teljes utószerkesztésnél végeztem

Például:

Eng.: For example, by looking for patterns in the co-occurrence of words and finding latent factors (e. g. , topics, frames, authors) that explain these patterns—at least mathematically.

DeepL: Például úgy, hogy mintákat keres a szavak együttes előfordulásában, és olyan látens tényezőket (pl. témák, keretek, szerzők) talál, amelyek ezeket a mintákat - legalábbis matematikailag - megmagyarázzák.

Részleges utószerkesztés: Például úgy, hogy mintákat keres a szavak együttes előfordulásában, és olyan látens tényezőket (pl. témák, keretek, szerzők) talál, amelyek ezeket a mintákat - legalábbis matematikailag - megmagyarázzák.

Teljes utószerkesztés: Ez például úgy történhet, hogy mintázatokat keres a szavak együttes előfordulásában, és olyan látens tényezőket (pl. témák, keretek, szerzők) talál, amelyek ezeket a mintákat - legalábbis matematikailag - megmagyarázzák.

9. A tartalom javítását mind a részleges, mind a teljes utószerkesztésnél el kellett végezni, igaz, ez nem volt gyakori.

Eng.: It has been used to study subjects such as media attention for political actors (Schuck, Xezonakis, Elenbaas, Banducci, & De Vreese, 2011; Vliegenthart, Boomgaarden, & Van Spanje, 2012) and framing in corporate news (Schultz, Kleinnijenhuis, Oegema, Utz, & Van Atteveldt, 2012).

DeepL: Olyan témák tanulmányozására használták, mint a politikai szereplőkre irányuló médiafigyelem (Schuck, Xezonakis, Elenbaas, Banducci és De Vreese, 2011; Vliegenthart, Boomgaarden és Van Spanje, 2012) és a vállalati hírek keretezése (Schultz, Kleinnijenhuis, Oegema, Utz és Van Atteveldt, 2012).

Teljes/részleges utószerkesztés: Olyan témák tanulmányozására használták, mint például a politikai szereplőkre irányuló médiafigyelem (Schuck, Xezonakis, Elenbaas, Banducci és De Vreese, 2011; Vliegenthart, Boomgaarden és Van Spanje, 2012), vagy a vállalati információk kontextusba foglalása (Schultz, Kleinnijenhuis, Oegema, Utz és Van Atteveldt, 2012).

## 6. Fordítási tapasztalatok összegzése

A gépi fordítás számos esetben, főleg a rövidebb, egyszerűbb mondatoknál gyakorlatilag kifogástalan eredményt ad. Nem nélkülözhető azonban a fordító ellenőrző munkája, mert bizonyos szavaknál és kifejezéseknél teljesen értelmetlen és hibás fordítási események is adódhatnak.

1.6 Fordítási tapasztalatok összegzése

A szöveg (Welbers et al., 2017)alapvető célja, hogy betekintést nyújtson az R programozási környezetben kifejlesztett szoftverek alkalmazására a szövegek elemzésében. Ez a terület a modern matematikai statisztika egészén belül is nagyon gyorsan fejlődik, szorosan kapcsolódva a számítógépes nyelv-elemzés, a többváltozós matematikai statisztikai és a mesterséges intelligencia kínálta lehetőségek bővüléséhez. A 2017-ben megjelent tanulmány átfogó képet kínál az R-ről, mint keretrendszerről, bevezetve az olvasót az R-alapú programok használatába, majd lépésről lépésre mutat be néhány számítógépes szövegelemzési alkalmazást. A témában készült publikációk közül ez az egyik legtöbbet idézett tanulmány: a scholar.google.com tudományos kereső összesen 353 hivatkozásról tud 2024.04.25-ig. A munka nem tévesztendő össze a később, más szerzők tollából megjelent azonos című könyvvel (Jockers & Thalken, 2020) A példák összeállítása módszertani szempontból megalapozott, de a szerzők több alkalommal használnak olyan rövidítéseket és utalásokat, melyek a témában nem járatos szakemberek számára nem értelmezhetőek számottevő előképzettség nélkül.

A tanulmány eredeti formájának visszaadásakor nehézséget jelentett a számítógépes outputok reprodukálása, ezért többször újra kellett szerkesztenem a táblázatokat.

# TEXT ANALYSIS IN R

Kasper Welbers[1], Wouter van Atteveldt[2] & Kenneth Benoit[3]

Institute for Media Studies, University of Leuven[1], Department of Communcation Science, VU University Amsterdam[2], Department of Methodology, London School of Economics and Political Science[3]

## Abstract

Computational text analysis has become an exciting research field with many applications in communication research. It can be a difficult method to apply, however, because it requires knowledge of various techniques, and the software required to perform most of these techniques is not readily available in common statistical software packages. In this teacher's corner, we address these barriers by providing an overview of general steps and operations in a computational text analysis project, and demonstrate how each step can be performed using the R statistical software. As a popular open-source platform, R has an extensive user community that develops and maintains a wide range of text analysis packages. We show that these packages make it easy to perform advanced text analytics.

With the increasing importance of computational text analysis in communication research (Boumans & Trilling, 2016; Grimmer & Stewart, 2013), many researchers face the challenge of learning how to use advanced software that enables this type of analysis. Currently, one of the most popular environments for computational methods and the emerging field of "data science" [1] is the R statistical software (R Core Team, 2017). However, for researchers that are not well-versed in programming, learning how to use R can be a challenge, and performing text analysis in particular can seem daunting. In this teacher's corner, we show that performing text analysis in R is not as hard as some might fear. We provide a step-bystep introduction into the use of common techniques, with the aim of helping researchers get acquainted with computational text analysis in general, as well as getting a start at performing advanced text analysis studies in R.

R is a free, open-source, cross-platform programming environment. In contrast to most programming languages, R was specifically designed for statistical analysis, which makes it highly suitable for data science applications. Although the learning curve for programming with R can be steep, especially for people without prior programming experience, the tools now available for carrying out text analysis in R make it easy to perform powerful, cutting-edge text analytics using only a few simple commands. One of the keys to R's explosive growth (Fox & Leanage, 2016; TIOBE, 2017) has been its densely populated collection of extension software libraries, known in R terminology as *packages*, supplied and maintained by R's extensive user community. Each package extends the functionality of the base R language and core packages, and in addition to functions and data must include documentation and examples, often in the form of vignettes demonstrating the use of the package. The best-known package repository, the Comprehensive R Archive Network (CRAN), currently has over 10,000 packages that are published, and which have gone through an extensive screening for procedural conformity and cross-platform compatibility before being accepted by the archive.[2] R thus features a wide range of inter-compatible packages, maintained and continuously updated by scholars, practitioners, and projects such as RStudio and rOpenSci. Furthermore, these packages may be installed easily and safely from within the R environment using a single command. R thus provides a solid bridge for developers and users of new analysis tools to meet, making it a very suitable programming environment for scientific collaboration.

Text analysis in particular has become well established in R. There is a vast collection of dedicated text processing and text analysis packages, from low-level string operations (Gagolewski, 2017) to advanced text modeling techniques such as fitting Latent Dirichlet Allocation models (Blei, Ng, & Jordan, 2003; Roberts et al., 2014)—nearly 50 packages in total at our last count. Furthermore, there is an increasing effort among developers to cooperate and coordinate, such as the rOpenSci special interest group.[3] One of the main advantages of performing text analysis in R is that it is often possible, and relatively easy, to switch between different packages or to combine them. Recent efforts among the R text analysis developers' community are designed to promote this interoperability to maximize flexibility and choice among users.[4] As a result, learning the basics for text analysis in R provides access to a wide range of advanced text analysis features.

**Structure of this Teacher's Corner**

This teacher's corner covers the most common steps for performing text analysis in R, from data preparation to analysis, and provides easy to replicate example code to perform each step. The example code is also digitally available in our online appendix, which is updated over time.[5] We focus primarily on bag-of-words text analysis approaches, meaning that only the frequencies of words per text are used and word positions are ignored. Although this drastically simplifies text content, research and many real-world applications show that word frequencies alone contain sufficient information for many types of analysis (Grimmer & Stewart, 2013).

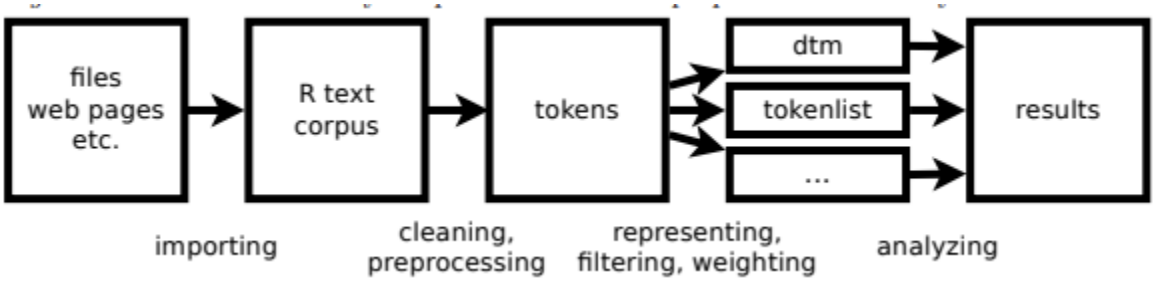
Table 1 presents an overview of the text analysis operations that we address, categorized in three sections. In the *data preparation* section we discuss five steps to prepare texts for analysis. The first step, *importing text*, covers the functions for reading texts from various types of file formats (e.g., txt, csv, pdf) into a *raw text corpus* in R. The steps *string operations* and *preprocessing* cover techniques for manipulating raw texts and processing them into *tokens* (i.e., units of text, such as words or word stems). The tokens are then used for creating the *document-term matrix* (DTM), which is a common format for representing a bag-of-words type corpus, that is used by many R text analysis packages. Other nonbag-of-words formats, such as the tokenlist, are briefly touched upon in the *advanced topics* section. Finally, it is a common step to *filter and weight* the terms in the DTM. These Table 1

*An overview of text analysis operations, with the R packages used in this teacher's corner*

| **Operation** | **R packages** | |
|---|---|---|
| | example | alternatives |
| **Data preparation** importing text | readtext | jsonlite, XML, antiword, readxl, pdftools |
| string operations | stringi | stringr |
| preprocessing | quanteda | stringi, tokenizers, snowballC, tm, etc. |
| document-term matrix (DTM) | quanteda | tm, tidytext, Matrix |
| filtering and weighting | quanteda | tm, tidytext, Matrix |
| **Analysis** dictionary | quanteda | tm, tidytext, koRpus, corpustools |

| | | |
|---|---|---|
| supervised machine learning | quanteda | RTextTools, kerasR, austin |
| unsupervised machine learning | topicmodels | quanteda, stm, austin, text2vec |
| text statistics | quanteda | koRpus, corpustools, textreuse |
| **Advanced topics** | | |
| advanced NLP | spacyr | coreNLP, cleanNLP, koRpus |
| word positions and syntax | corpustools | quanteda, tidytext, koRpus |

*Figure 1*. Order of text analysis operations for data preparation and analysis.




Steps are generally performed in the presented sequential order (see Figure 1 for conceptual illustration). As we will show, there are R packages that provide convenient functions that manage multiple data preparation steps in a single line of code. Still, we first discuss and demonstrate each step separately to provide a basic understanding of the purpose of each step, the choices that can be made and the pitfalls to watch out for.

The *analysis* section discusses four text analysis methods that have become popular in communication research (Boumans & Trilling, 2016) and that can be performed with a DTM as input. Rather than being competing approaches, these methods have different advantages and disadvantages, so choosing the best method for a study depends largely on the research question, and sometimes different methods can be used complementarily (Grimmer & Stewart, 2013). Accordingly, our recommendation is to become familiar with each type of method. To demonstrate the general idea of each type of method, we provide code for typical analysis examples. Furthermore, It is important to note that different types of analysis can also have different implications for how the data should be prepared. For each type of analysis we therefore address general considerations for data preparation.

Finally, the additional *advanced topics* section discusses alternatives for data prepara-

tion and analysis that require external software modules or that go beyond the bag-of-words assumption, using word positions and syntactic relations. The purpose of this section is to provide a glimpse of alternatives that are possible in R, but might be more difficult to use.

Within each category we distinguish several groups of operations, and for each operation we demonstrate how they can be implemented in R. To provide parsimonious and easy to replicate examples, we have chosen a specific selection of packages that are easy to use and broadly applicable. However, there are many alternative packages in R that can perform the same or similar operations. Due to the open-source nature of R, different people from often different disciplines have worked on similar problems, creating

some duplication in functionality across different packages. This also offers a range of choice, however, providing alternatives to suit a user's needs and tastes. Depending on the research project, as well as personal preference, other packages might be better suited to different readers. While a fully comprehensive review and comparison of text analysis packages for R is beyond our scope here—especially given that existing and new packages are constantly being developed—we have tried to cover, or at least mention, a variety of alternative packages for each text analysis operation.[6] In general, these packages often use the same standards for data formats, and thus are easy to substitute or combine with the other packages discussed in this teacher's corner.

**Data preparation**

Data preparation is the starting point for any data analysis. Not only is computational text analysis no different in this regard, but also frequently presents special challenges for data preparation that can be daunting for novice and advanced practitioners alike. Furthermore, preparing texts for analysis requires making choices that can affect the accuracy, validity, and findings of a text analysis study as much as the techniques used for the analysis (Crone, Lessmann, & Stahlbock, 2006; Günther & Quandt, 2016; Leopold & Kindermann, 2002). Here we distinguish five general steps: importing text, string operations, preprocessing, creating a document-term matrix (DTM), and filtering and weighting the DTM.

**Importing text**

Getting text into R is the first step in any R-based text analytic project. Textual data can be stored in a wide variety of file formats. R natively supports reading regular flat text files such as CSV and TXT, but additional packages are required for processing formatted text files such as JSON (Ooms, 2014), HTML, and XML (Lang & the CRAN Team, 2017), and for reading complex file formats such as Word (Ooms, 2017a), Excel (Wickham & Bryan, 2017) and PDF (Ooms, 2017b). Working with these different packages and their different interfaces and output can be challenging, especially if different file formats are used together in the same project. A convenient solution for this problem is the readtext package(Benoit & Obeng, 2017), that wraps various import packages together to offer a single catch-all function for importing many types of data in a uniform format. The following lines of code illustrate how to read a CSV file with the readtext function, by providing the path to the file as the main argument (the path can also be an URL, as used in our example). An online appendix with copyable code available from https://github.com/kasperwelbers/ text_analysis_in_R.

```
install.packages("readtext")

library(readtext)

# url to Inaugural Address demo data that is provided by the readtext package filepath <-
"http://bit.ly/2uhqjJE?.csv"

rt <- readtext(filepath, text_field = "texts") rt
```

```
readtext object consisting of 5 documents and 3 docvars.
# data.frame [5 x 5]
doc_id  text       Year President  FirstName
        <chr> <chr>     <int>    <chr>   <chr>
  1     2uhqjJE?.csv.1 "\"Fellow-Cit\"..."1789   Washington      George
  2     2uhqjJE?.csv.2 "\"Fellow cit\"...  "1793   Washington      George
  3     2uhqjJE?.csv.3 "\"When it wa\"..."1797   Adams           John
  4     2uhqjJE?.csv.4 "\"Friends an\"... "1801   Jefferson       Thomas
  5     2uhqjJE?.csv.5 "\"Proceeding\"..."1805   Jefferson       Thomas
```

The same function can be used for importing all formats mentioned above, and the path can also reference a (zip) folder to read all files within. In most cases, the only thing that has to be specified is the name of the field that contains the texts. Not only can multiple files be references using simple, "glob"-style pattern matches, such as /myfiles/*.txt, but also the same command will recurse through sub-directories to locate these files. Each file is automatically imported according to its format, making it very easy to import and work with data from different input file types.

Another important consideration is that texts can be represented with different character encodings. Digital text requires binary code to be mapped to semantically meaningful characters, but many different such mappings exist, with widely different methods of encoding "extended" characters, including letters with diacritical marks, special symbols, and emoji. In order to be able to map all known characters to a single scheme, the Unicode standard was proposed, although it also requires a digital encoding format (such as the UTF-8 format, but also UTF-16 or UTF-32). Our recommendation is simple: in R, ensure that all texts are encoded as UTF-8, either by reading in UTF-8 texts, or converting them from a known encoding upon import. If the encoding is unknown, readtext's encoding function can be used to guess the encoding. readtext can convert most known encodings (such as ISO-8859-2 for Central and Eastern European languages, or Windows-1250 for Cyrillic—although there are hundreds of others) into the common UTF-8 standard. R also offers additional low-level tools for converting character encodings, such as a bundled version of the GNU libiconv library, or conversion though the stringi package.

**String operations**

One of the core requirements of a framework for computational text analysis is the ability to manipulate digital texts. Digital text is represented as a sequence of characters, called a string. In R, strings are represented as objects called "character" types, which are vectors of strings. The group of string operations refers to the low-level operations for working with textual data. The most common string operations are joining, splitting, and extracting parts of strings (collectively referred to as *parsing*) and the use of *regular expressions* to find or replace patterns.

Although R has numerous built-in functions for working with character objects, we recommend using the stringi package (Gagolewski, 2017) instead. Most importantly, because stringi uses the International Components for Unicode (ICU) library for proper Unicode support, such as implementing Unicode character categories (such as punctuation or spacing) and Unicode-defined rules for case conversion that

work correctly in all languages. An alternative is the stringr package (Wickham, 2017), which uses stringi as a backend, but has a simpler syntax that many end users will find sufficient for their needs.

It is often unnecessary to perform manual, low-level string operations, because the most important applications of string operations for text analysis are built into common text analysis packages. Nevertheless, access to low-level string operations provides a great deal of versatility, which can be crucial when standardized solutions are not an option for a specific use case. The following example shows how to perform some basic cleaning with stringi functions: removing boilerplate content in the form of markup tags, stripping extraneous whitespace, and converting to lower case.

```
[1] "the first string" "the second string"
```

As with most functions in R, stringi operations are *vectorized*, meaning they apply to each element of a vector. Manipulation of vectors of strings is the recommended approach in R, since looping over each element and processing it separately in R is very inefficient.

**Preprocessing**

```
install.packages("stringi")

library(stringi)

x <- c("The first string", ’ The <font size="6">second
string</font>’)

x <- stri_replace_all(x, "", regex = "<.*?>")                    # remove html tags

x <- stri_trim(x)                                                # strip surrounding

whitespace x <- stri_trans_tolower(x)                            # transform to lower case x
```

For most computational text analysis methods, full texts must be *tokenized* into smaller, more specific text features, such as words or word combinations. Also, the computational performance and accuracy of many text analysis techniques can be improved by normalizing features, or by removing “stopwords”: words designated in advance to be of no interest, and which are therefore discarded prior to analysis. Taken together, these preparatory steps are commonly referred to as “preprocessing”. Here we first discuss several of the most common preprocessing techniques, and show how to perform each technique with the quanteda package.

In practice, all of these preprocessing techniques can be applied in one function when creating a document-term matrix, as we will demonstrate in the *DTM* section. Here, we show each step separately to illustrate what each technique does.

**Tokenization.** Tokenization is the process of splitting a text into tokens. This is crucial for computational text analysis, because full texts are too specific to perform any meaningful computations with. Most often tokens are words, because these are the most common semantically meaningful components of texts.

For many languages, splitting texts by words can mostly be done with low-level string processing due to clear indicators of word boundaries, such as white spaces, dots and commas. A good tokenizer, however, must also be able to handle certain exceptions, such as the period in the title “Dr.”, which can be confused for a sentence boundary. Furthermore, tokenization is more difficult for languages where words are not clearly separated by white spaces, such as Chinese and Japanese. To deal with these cases, some tokenizers include dictionaries of patterns for splitting texts. In R, the stringi package is often used for sentence and word disambiguation, for which it leverages dictionaries from the ICU library. There is also a dedicated package for text tokenization, called tokenizers (Mullen, 2016b).

The following code uses quanteda’s (Benoit et al., 2017) tokens function to split a single sentence into words. The tokens function returns a list whose elements each contain the tokens of the input texts as a character vector.

```
install.packages("quanteda")

library(quanteda)

text <- "An example of preprocessing techniques" toks <-
tokens(text) # tokenize into unigrams toks
```

```
tokens from 1 document. text1 :

[1] "An"          "example"       "of" “preprocessing”

[5] "techniques"
```

**Normalization: Lowercasing and stemming.** The process of normalization broadly refers to the transformation of words into a more uniform form. This can be important if for a certain analysis a computer has to recognize when two words have (roughly) the same meaning, even if they are written slightly differently. Another advantage is that it reduces the size of the vocabulary (i.e., the full range of features used in the analysis). A simple but important normalization techniques is to make all text lower case. If we do not perform this transformation, then a computer will not recognize that two words are identical if one of them was capitalized because it occurred at the start of a sentence.

Another argument for normalization is that a base word might have different morphological variations, such as the suffixes from conjugating a verb, or making a noun plural. For purposes of analysis, we might wish to consider these variations as equivalent because of their close semantic relation, and because reducing the feature space is generally desirable when multiple features are in fact closely related. A technique for achieving this is *stemming*, which is essentially a rule-based algorithm that converts inflected forms of words into their base forms (stems). A more advanced technique is *lemmatization*, which uses a dictionary to replace words with their morphological root form. However, lemmatization in R requires external software modules (see the *advanced preprocessing* section for instructions) and for weakly inflected languages such as modern English, stemming is often sufficient. In R, the SnowballC package (Bouchet-Valat, 2014; Porter, 2001) is used in many text analysis packages (such as quanteda and tm) to implement stemming, and currently supports 15 different languages. Lowercasing and stemming of character, tokens, or feature vectors can be performed in quanteda with the _tolower and _wordstem functions, such as char_tolower to convert character objects to lower case, or tokens_wordstem to stem tokens.

```
toks <- tokens_tolower(toks) toks <-
tokens_wordstem(toks) toks
```

```
tokens from 1 document. text1 :
[1] "an" "exampl"          "of"


"preprocess" "techniqu"
```

In this example we see that the difference between “an” and “An” is eliminated due to lowercasing. The words “example” and “techniques” are reduced to “exampl” and “techniqu”, such that any distinction between singular and plural forms is removed.

**Removing stopwords.** Common words such as “the” in the English language are rarely informative about the content of a text. Filtering these words out has the benefit of reducing the size of the data, reducing computational load, and in some cases also improving accuracy. To remove these words beforehand, they are matched to predefined lists of “stop words” and deleted. Several text analysis packages provide stopword lists for various languages, that can be used to manually filter out stopwords. In quanteda, the stopwords function returns a character vector of stopwords for a given language. A total of 17 languages are currently supported.

```
sw <- stopwords("english")          # get character vector of stopwords
head(sw)                            # show head (first 6) stopwords
[1] "i"    "me"     "my"                    "myself" "we"    "our"
tokens_remove(toks, sw)
tokens from 1 document.
text1 :
[1] "exampl"     "preprocess" "techniqu"
```

Care should be taken to perform some preprocessing steps in the correct order, for instance removing stopwords prior to stemming, otherwise “during” will be stemmed into “dure” and not matched to a stopword “during”. Case conversion may also create sequencing issues, although the default stopword matching used by quanteda is case-insensitive.

Conveniently, the preprocessing steps discussed above can all be performed with a single function that will automatically apply the correct order of operations. We will demonstrate this in the next section.

**Document-term matrix**

The document term matrix (DTM) is one of the most common formats for representing a text corpus (i.e. a collection of texts) in a bag-of-words format. A DTM is a matrix in which rows are documents, columns are terms, and cells indicate how often each term occurred in each document. The advantage of this representation is that it allows the data to be analyzed with vector and matrix algebra, effectively moving from text to numbers. Furthermore, with the use of special matrix formats for sparse matrices, text data in a DTM format is very memory efficient and can be analyzed with highly optimized operations.

Two of the most established text analysis packages in R that provide dedicated DTM classes are tm and quanteda. Of the two, the venerable tm package is the more commonly used, with a user base of almost 10 years (Meyer, Hornik, & Feinerer, 2008) and several other R packages using its DTM classes (*DocumentTermMatrix* and *TermDocumentMatrix*) as inputs for their analytic functions. The quanteda package is a more recently developed package, built by a team supported by an ERC grant to provide state-of-the-art, high performance text analysis. Its sparse DTM class, known as a dfm or *document-feature matrix*, is based on the powerful Matrix package (Bates & Maechler, 2015) as a backend, but includes functions to convert to nearly every other sparse document-term matrix used in other R packages (including the tm formats). The performance and flexibility of quanteda's dfm format lends us to recommend it over the tm equivalent.

Another notable alternative is the tidytext package (Silge & Robinson, 2016). This is a text analysis package that is part of the *Tidyverse*[7]—a collection of R packages with a common philosophy and format. Central to the Tidyverse philosophy is that all data is arranged as a table, where (1) "each variable forms a column", (2) "each observation forms a row", and (3) "each type of observational unit forms a table" (Wickham et al., 2014, 4). As such, tidytext does not strictly use a document term matrix, but instead represents the same data in a long format, where each (non-zero) value of the DTM is a row with the columns document, term, and count (note that this is essentially a triplet format for sparse matrices, with the columns specifying the row, column and value). This format can be less memory efficient and make matrix algebra less easily applicable, but has the advantage of being able to add more variables (e.g., a sentiment score) and enables the use of the entire Tidyverse arsenal. Thus, for users that prefer the tidy data philosophy, tidytext can be a good alternative package to quanteda or tm, although these packages can also be used together quite nicely depending on the particular operations desired.

Consistent with the other examples in this teacher's corner, we demonstrate the creation of DTMs using the quanteda package. Its dfm function provides a single line solution for creating a DTM from raw text, that also integrates the preprocessing techniques discussed above. These may also be built up through a sequence of lower-level functions, but many users find it convenient to go straight from a text or corpus to a DTM using this single function.

```
text <- c(d1 = "An example of preprocessing techniques",
d2 = "An additional example",
d3 = "A third example")
dtm <- dfm(text,                    # input text
tolower = TRUE, stem = TRUE,# set lowercasing and stemming to TRUE remove = stopwords("english"))
                    # provide the stopwords for deletion
dtm
```

```
Document-feature matrix of: 3 documents, 5 features (53.3% sparse).
3 x 5 sparse Matrix of class "dfmSparse"
features
docs exampl preprocess techniqu addit third
 d1      1          1          1          0          0
```

```
d2  1  0  0  1  0
d3  1  0  0  0  1
dtm <- dfm(text,  # input text
tolower = TRUE, stem = TRUE,# set lowercasing and stemming to TRUE remove = stopwords("english"))
# provide the stopwords for deletion
```

The DTM can also be created from a quanteda corpus object, which stores text and associated meta-data, including document-level variables. When a corpus is tokenized or converted into a DTM, these document-level variables are saved in the object, which can be very useful later when the documents in the DTM need to be used as covariates in supervised machine learning. The stored document variables also make it possible to aggregate quanteda objects by groups, which is extremely useful when texts are stored in small units—like Tweets—but need to be aggregated in a DTM by grouping variables such as users, dates, or combinations of these.

Because quanteda is compatible with the pkgreadtext package, creating a corpus from texts on disk takes only a single additional step. In the following example we create a DTM from the readtext data as imported above.

```
fulltext <- corpus(rt)  # create quanteda corpus
dtm <- dfm(fulltext, tolower = TRUE, stem = TRUE,  # create dtm with preprocessing
remove_punct = TRUE,remove = stopwords("english"))
dtm
```

```
Document-feature matrix of: 5 documents, 1,405 features (67.9% sparse).
```

**Filtering and weighting**

Not all terms are equally informative for text analysis. One way to deal with this is to remove these terms from the DTM. We have already discussed the use of stopword lists to remove very common terms, but there are likely still other common words and this will be different between corpora. Furthermore, it can be useful to remove very rare terms for tasks such as category prediction (Yang & Pedersen, 1997) or topic modeling (Griffiths & Steyvers, 2004). This is especially useful for improving efficiency, because it can greatly reduce the size of the vocabulary (i.e., the number of unique terms), but it can also improve accuracy. A simple but effective method is to filter on document frequencies (the number of documents in which a term occurs), using a threshold for minimum and maximum number (or proportion) of documents (Griffiths & Steyvers, 2004; Yang & Pedersen, 1997).

Instead of removing less informative terms, an alternative approach is assign them variable weights. Many text analysis techniques perform better if terms are weighted to take an estimated information value into account, rather than directly using their occurrence frequency. Given a sufficiently large corpus, we can use information about the distribution of terms in the corpus to estimate this information value. A popular

weighting scheme that does so is term frequency-inverse document frequency (*tf-idf*), which down-weights that occur in many documents in the corpus.

Using a document frequency threshold and weighting can easily be performed on a DTM. quanteda includes the functions docfreq, tf, and tfidf, for obtaining document frequency, term frequency, and *tf-idf* respectively. Each function has numerous options for implementing the SMART weighting scheme (Manning et al., 2008). As a high-level wrapper to these, quanteda also provides the dfm_weight function. In the example below, the word "senat[e]" has a higher weight than the less informative term "among", which both occur once in the first document.

```
doc_freq <- docfreq(dtm)          # document frequency per term (column)
dtm <- dtm[, doc_freq >= 2]       # select terms with doc_freq >= 2 dtm <- dfm_weight(dtm,
"tfidf")                          # weight the features using tf-idf head(dtm)
```

Document-feature matrix of: 5 documents, 6 features (40% sparse).

5 x 6 sparse Matrix of class "dfmSparse" features

| docs | fellow-citizen | senat | hous | repres | among | life |
|---|---|---|---|---|---|---|
| text1 | 0.2218487 | 0.39794 | 0.79588 | 0.4436975 | 0.09691001 | 0.09691001 |
| text2 | 0 | 0 | 0 | 0 | 0 | 0 |
| text3 | 0.6655462 | 0.39794 | 1.19382 | 0.6655462 | 0.38764005 | 0.19382003 |
| text4 | 0.4436975 | 0 | 0 | 0.2218487 | 0.09691001 | 0.09691001 |
| text5 | 0 | 0 | 0 | 0 | 0.67837009 | 0.19382003 |

**Analysis**

For an overview of text analysis approaches we build on the classification proposed by Boumans and Trilling (2016) in which three approaches are distinguished: *counting and dictionary* methods, *supervised machine learning*, and *unsupervised machine learning*. They position these approaches, in this order, on a dimension from most deductive to most inductive. Deductive, in this scenario, refers to the use of an *a priori* defined coding scheme. In other words, the researchers know beforehand what they are looking for, and only seek to automate this analysis. The relation to the concept of deductive reasoning is that the researcher assumes that certain rules, or premises, are true (e.g., a list of words that indicates positive sentiment) and thus can be applied to draw conclusions about texts. Inductive, in contrast, means here that instead of using an a priori coding scheme, the computer algorithm itself somehow extracts meaningful codes from texts. For example, by looking for patterns in the co-occurrence of words and finding latent factors (e.g., topics, frames, authors) that explain these patterns—at least mathematically. In terms of inductive reasoning, it can be said that the algorithm creates broad generalizations based on specific observations.

In addition to these three categories, we also consider a *statistics* category, encompassing all techniques for describing a text or corpus in numbers. Like unsupervised learning, these techniques are inductive in the sense that no a priori coding scheme is used, but they do not use machine learning.

For the example code for each type of analysis, we use the Inaugural Addresses of US presidents (N = 58) that is included in the quanteda package.

```
dtm <- dfm(data_corpus_inaugural, stem = TRUE, remove = stopwords("english"),
remove_punct = TRUE)

dtm
```

```
## Document-feature matrix of: 58 documents, 5,405 features (89.2% sparse).
```

**Counting and dictionary**

The dictionary approach broadly refers to the use of patterns—from simple keywords to complex Boolean queries and regular expressions—to count how often certain concepts occur in texts. This is a deductive approach, because the dictionary defines a priori what codes are measured and how, and this is not affected by the data.[8] Using dictionaries is a computationally simple but powerful approach. It has been used to study subjects such as media attention for political actors (Schuck, Xezonakis, Elenbaas, Banducci, & De Vreese, 2011; Vliegenthart, Boomgaarden, & Van Spanje, 2012) and framing in corporate news (Schultz, Kleinnijenhuis, Oegema, Utz, & Van Atteveldt, 2012). Dictionaries are also a popular approach for measuring sentiment (De Smedt & Daelemans, 2012; Mostafa, 2013; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) as well as other dimensions of subjective language (Tausczik & Pennebaker, 2010). By combining this type of analysis with information from advanced NLP techniques for identifying syntactic clauses, it also becomes possible to perform more fine-grained analyses, such as sentiment expressions attributed to specific actors (Van Atteveldt, 2008), or actions and affections from one actor directed to another (Van Atteveldt, Sheafer, Shenhav, & Fogel-Dror, 2017).

The following example shows how to apply a dictionary to a quanteda DTM. The first step is to create a dictionary object (here called myDict), using the dictionary function. For simplicity our example uses a very simple dictionary, but it is is also possible to import large, pre-made dictionaries, including files in other text analysis dictionary formats such as LIWC, Wordstat, and Lexicoder. Dictionaries can also be written and imported from YAML files, and can include patterns of fixed matches, regular expressions, or the simpler “glob” pattern match (using just * and ? for wildcard characters) common in many dictionary formats. With the dfm_lookup function, the dictionary object can then be applied on a DTM to create a new DTM in which columns represent the dictionary codes.

```
myDict <- dictionary(list(terror = c("terror*"), economy = c("job*", "business*", "econom*")))

dict_dtm <- dfm_lookup(dtm, myDict, nomatch = "_unmatched") tail(dict_dtm)
```

```
Document-feature matrix of: 6 documents, 3 features (16.7% sparse).

6 x 3 sparse Matrix of class "dfmSparse" features
```

| docs | terror | economy | _unmatched |
|---|---|---|---|
| 1997-Clinton | 2 | 3 | 1125 |
| 2001-Bush | 0 | 2 | 782 |
| 2005-Bush | 0 | 1 | 1040 |

| | | | |
|---|---|---|---|
| 2009-Obama | 1 | 7 | 1165 |
| 2013-Obama | 0 | 6 | 1030 |
| 2017-Trump | 1 | 5 | 709 |

**Supervised machine learning**

The supervised machine learning approach refers to all classes of techniques in which an algorithm learns patterns from an annotated set of training data. The intuitive idea is that these algorithms can learn how to code texts if we give them enough examples of how they should be coded. A straightforward example is sentiment analysis, using a set of texts that are manually coded as *positive*, *neutral*, or *negative*, based on which the algorithm can learn which features (words or word combinations) are more likely to occur in positive or negative texts. Given an unseen text (from which the algorithm was not trained), the sentiment of the text can then be estimated based on the presence of these features. The deductive part is that the researchers provide the training data, which contains good examples representing the categories that the researchers are attempting to predict or measure. However, the researchers do not provide explicit rules for how to look for these codes. The inductive part is that the supervised machine learning algorithm learns these rules from the training data. To paraphrase a classic syllogism: if the training data is a list of people that are either mortal or immortal, then the algorithm will learn that all men are extremely likely to be mortal, and thus would estimate that Socrates is mortal as well.

To demonstrate this, we train a model to predict whether an Inaugural Address was given before World War II—which we expect because prominent issues shift over time, and after wars in particular. Some dedicated packages for supervised machine learning are RTextTools (Jurka, Collingwood, Boydstun, Grossman, & Van Atteveldt, 2014) and kerasR (Arnold, 2017b). For this example, however, we use a classifier that is included in quanteda. Before we start, we set a custom seed for R's random number generator so that the results of the random parts of the code are always the same. To prepare the data, we add the document (meta) variable is_prewar to the DTM that indicates which documents predate 1945. This is the variable that our model will try to predict. We then split the DTM into training (train_dtm) and test (test_dtm) data, using a random sample of 40 documents for training and the remaining 18 documents for testing. The training data is used to train a multinomial Naive Bayes classifier (Manning et al., 2008, Ch. 13) which we assign to nb_model. To test how well this model predicts whether an Inaugural Address predates the war, we predict the code for the test data, and make a table in which the rows show the prediction and the columns show the actual value of the is_prewar variable.

```
set.seed(2)
# create a document variable indicating pre or post war
docvars(dtm, "is_prewar") <- docvars(dtm, "Year") < 1945
# sample 40 documents for the training set and use remaining (18) for testing
train_dtm <- dfm_sample(dtm, size = 40)
test_dtm <- dtm[setdiff(docnames(dtm), docnames(train_dtm)), ]
# fit a Naive Bayes multinomial model and use it to predict the test data
 nb_model <- textmodel_NB(train_dtm, y = docvars(train_dtm, "is_prewar"))
 pred_nb <- predict(nb_model, newdata = test_dtm)
# compare prediction (rows) and actual is_prewar value (columns) in a table
table(prediction = pred_nb$nb.predicted, is_prewar = docvars(test_dtm, "is_prewar"))
```

```
is_prewar
prediction FALSE TRUE
         FALSE 8     0
         TRUE  0     10
```

The results show that the predictions are perfect. Of the eight times that FALSE (i.e. Inaugural Address does not predate the war) was predicted, and the 10 times that TRUE was predicted, this was actually the case.

**Unsupervised machine learning**

In unsupervised machine learning approaches, no coding rules are specified and no annotated training data is provided. Instead, an algorithm comes up with a model by identifying certain patterns in text. The only influence of the researcher is the specification of certain parameters, such as the number of categories into which documents are classified. Popular examples are topic modeling for automatically classifying documents based on an underlying topical structure (Blei et al., 2003; Roberts et al., 2014) and the "Wordfish" parametric factor model (Proksch & Slapin, 2009) for scaling documents on a single underlying dimension, such as left-right ideology.

Grimmer and Stewart (2013) argue that supervised and unsupervised machine learning are not competitor methods, but fulfill different purposes and can very well be used to complement each other. Supervised methods are the most suitable approach if documents need to be placed in predetermined categories, because it is unlikely that an unsupervised method will yield a categorization that reflects these categories and how the researcher interprets them. The advantage of the somewhat unpredictable nature of unsupervised methods is that it can come up with categories that the researchers had not considered. (Conversely, this may also present challenges for post-hoc interpretation when results are unclear.)

To demonstrate the essence of unsupervised learning, the example below shows how to fit a topic model in R using the topicmodels package (Grun & Hornik, 2011). To focus more specifically on topics within the inaugural addresses, and to increase the number of texts to model, we first split the texts by paragraph and

create a new DTM. From this DTM we remove terms with a document frequency of five and lower to reduce the size of the vocabulary (less important for current example) and use quanteda's convert function to convert the DTM to the format used by topicmodels. We then train a vanilla LDA topic model(Blei et al., 2003) with five topics—using a fixed seed to make the results reproducible, since LDA is non-deterministic.

```
install.packages("topicmodels") library(topicmodels) texts =
corpus_reshape(data_corpus_inaugural, to = "paragraphs")

par_dtm <- dfm(texts, stem = TRUE,      # create a document-term matrix

remove_punct = TRUE, remove = stopwords("english"))

par_dtm <- dfm_trim(par_dtm, min_count = 5)    # remove rare terms

par_dtm <- convert(par_dtm, to = "topicmodels") # convert to topicmodels format

set.seed(1)

lda_model <- topicmodels::LDA(par_dtm, method = "Gibbs", k = 5) terms(lda_model, 5)
```

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|---|
| [1,] | "govern" | "nation" | "great" | "us" | "shall" |
| [2,] | "state" | "can" | "war" | "world" | "citizen" |
| [3,] | "power" | "must" | "secur" | "new" | "peopl" |
| [4,] | "constitut" | "peopl" | "countri" | "american" | "duti" |
| [5,] | "law" | "everi" | "unit" | "america" | "countri" |

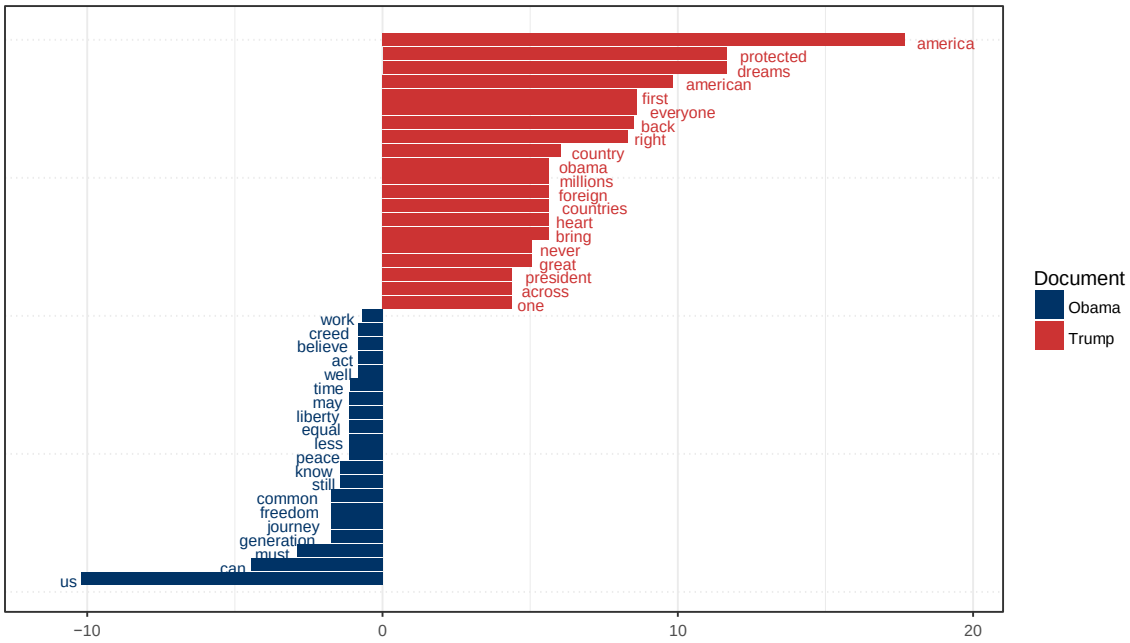
The results show the first five terms of the five topics. Although this is a basic example, the idea of "topics" being found bottom-up from the data can be seen in the semantic coherence of terms within the same topic. In particular, topic one seems to revolve around governance, with the terms "govern[ance]", "power", "state", "constitut[ion]", and "law".

**Statistics**

Various statistics can be used to describe, explore and analyze a text corpus. An example of a popular technique is to rank the information value of words inside a corpus and then visualize the most informative words as a word cloud to get a quick indication of what a corpus is about. Text statistics (e.g., average word and sentence length, word and syllable counts) are also commonly used as an operationalization of concepts such as readability (Flesch, 1948) or lexical diversity (McCarthy & Jarvis, 2010). A wide range of such measures is available in R with the koRpus package (Michalke, 2017). Furthermore, there are many useful applications of calculating term and document similarities (which are often based on the inner product of a DTM or transposed DTM), such as analyzing semantic relations between words or concepts and measuring content homogeneity. Both techniques are supported in quanteda, corpustools (Welbers & Van Atteveldt, 2016), or dedicated packages such as textreuse (Mullen, 2016a) for text overlap.

A particularly useful technique is to compare the term frequencies of two corpora, or between two subsets of the same corpus. For instance, to see which words are more likely to occur in documents about a certain topic. In addition to providing a way to quickly explore how this topic is discussed in the corpus, this can provide input for developing better queries. In the following example we show how to perform this technique in quanteda (results in Figure 2).

*Figure 2*. Keyness plot comparing relative word frequencies for Trump and Obama.




chi2

```
# create DTM that contains Trump and Obama speeches
corpus_pres = corpus_subset(data_corpus_inaugural,
President %in% c("Obama", "Trump"))
dtm_pres = dfm(corpus_pres, groups = "President",
remove = stopwords("english"), remove_punct = TRUE)
# compare target (in this case Trump) to rest of DTM (in this case only Obama).
keyness = textstat_keyness(dtm_pres, target = "Trump")
textplot_keyness(keyness)
```

output in Figure 2.

Here, the signed $\chi^2$ measure of association indicates that “america”, “american”, and “first” were used with far greater frequency by Trump than Obama, while “us”, “can”, “freedom”, “peace”, and “liberty” were among the words much more likely to be used by Obama than by Trump.

**Advanced topics**

The data preparation and bag-of-words analysis techniques discussed above are the basis for the majority of the text analysis approaches that are currently used in communication research. For certain types of analyses, however, techniques might be required that rely on external software modules, are more computationally demanding, or that are more complicated to use. In this section we briefly elaborate on some of these advanced approaches that are worth taking note of.

**Advanced NLP**

In addition to the preprocessing techniques discussed in the data preparation section, there are powerful preprocessing techniques that rely on more advanced natural language processing (NLP). At present, these techniques are not available in native R, but rely on external software modules that often have to be installed outside of R. Many of the advanced NLP techniques are also much more computationally demanding, thus taking more time to perform. Another complication is that these techniques are language specific, and often only available for English and a few other big languages.

Several R packages provide interfaces for external NLP modules, so that once these modules have been installed, they can easily be used from within R. The coreNLP package (Arnold & Tilton, 2016) provides bindings for Stanford CoreNLP java library (Manning et al., 2014), which is a full NLP parser for English, that also supports (albeit with limitations) Arabic, Chinese, French, German, and Spanish. The spacyr package (Benoit & Matsuo, 2017) provides an interface for the spaCy module for Python, which is comparable to CoreNLP but is faster, and supports English and (again with some limitations) German and French. A third package, cleanNLP (Arnold, 2017a), conveniently wraps both CoreNLP and spaCy, and also includes a minimal back-end that does not rely on external dependencies. This way it can be used as a swiss army knife, choosing the approach that best suits the occasion and for which the back-end is available, but with standardized output and methods.

Advanced NLP parsers generally perform all techniques in one go. In the following example we use the spacyr package to parse a sentence, that will be used to illustrate four advanced NLP techniques: lemmatization, part-of-speech (POS) tagging, named entity recognition (NER) and dependency parsing.

```
install.packages("spacyr") library(spacyr) spacy_initialize()

d <- spacy_parse("Bob Smith gave Alice his login information.", dependency = TRUE)

d[, -c(1,2)]
```

| token_id | token | lemma | pos | head_token_id | dep_rel | entity |
|---|---|---|---|---|---|---|
| 1 | Bob | bob | PROPN | 2 | compound | PERSON_B |
| 2 | Smith | smith | PROPN | 3 | nsubj | PERSON_I |
| 3 | gave | give | VERB | 3 | ROOT | |
| 4 | Alice | alice | PROPN | 3 | dative | PERSON_B |
| 5 | his | -PRON- | ADJ | 7 | poss | |
| 6 | login | login | NOUN | 7 | compound | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 7 | information | information | NOUN | 3 | dobj | |
| 8 | . | . | PUNCT | | 3 | punct |

**Lemmatization.** Lemmatization fulfills a similar purpose as stemming, but instead of cutting off the ends of terms to normalize them, a dictionary is used to replace terms with their lemma. The main advantage of this approach is that it can more accurately normalize different verb forms—such as "gave" and "give" in the example—which is particularly important for heavily inflected languages such as Dutch or German.

**Part-of-speech tagging.** POS tags are morpho-syntactic categories for words, such as nouns, verbs, articles and adjectives. In the example we see three proper nouns (PROPN), a verb (VERB) an adjective (ADJ), two nouns (NOUN), and punctuation (PUNCT). This information can be used to focus an analysis on certain types of grammar categories, for example, using nouns and proper names to measure similar events in news items (Welbers, Van Atteveldt, Kleinnijenhuis, & Ruigrok, 2016), or using adjectives to focus on subjective language (De Smedt & Daelemans, 2012). Similarly, it is a good approach for filtering out certain types of words, such as articles or pronouns.

**Named entity recognition.** Named entity recognition is a technique for identifying whether a word or sequence of words represents an entity and what type of entity, such as a person or organization. Both "Bob Smith" and "Alice" are recognized as persons. Ideally, named entity recognition is paired with co-reference resolution. This is a technique for grouping different references to the same entity, such as anaphora (e.g., he, she, the president). In the example sentence, the word "his" refers to "Bob Smith". Co-reference resolution is currently only supported by Stanford CoreNLP, but discussion on the spaCy GitHub page suggests that this feature is on the agenda.

**Dependency parsing.** Dependency parsing provides the syntactic relations between tokens, which can be used to analyze texts at the level of syntactic clauses (Van Atteveldt, 2008). In the spacyr output this information is given in the head_token_i and dep_rel columns, where the former indicates to what token a token is related and the latter indicates the type of relation. For example, we see that "Bob" is related to "Smith" (head_token_i 2) as a compound, thus recognizing "Bob Smith" as a single entity. Also, since "Smith" is the nominal subject (nsubj) of the verb "gave", and Alice is the dative case (dative) we know that "Bob Smith" is the one who gives to "Alice". This type of information can for instance be used to analyze who is attacking whom in news coverage about the Gaza war (Van Atteveldt et al., 2017).

### Word positions and syntax

As discussed above, the bag-of-words representation of texts is memory-efficient and convenient for various types of analyses, and this often outweighs the disadvantage of losing information by dropping the word positions. For some analyses, however, the order of words and syntactical properties can be highly beneficial if not crucial. In this section we address some text representations and analysis techniques where word positions are maintained.

A simple but potentially powerful solution is to use higher order n-grams. That is, instead of tokenizing texts into single words (n = 1; *unigrams*), sequences of two words (n = 2; *bigrams*), three words (n = 3; *trigrams*) or more are used.[9] The use of higher order ngrams is often optional in tokenization functions. quanteda makes it possible to form n-grams when tokenizing, or to form n-grams from tokens already formed. Other options include the formation of "skip-grams", or n-grams from words with variable windows of adjacency. Such non-adjacent collocations form the basis for counting weighted proximity vectors, used in vector-space network-based models built on deep learning techniques (Mikolov, Chen,

Corrado, & Dean, 2013; Selivanov, 2016). Below, we illustrate how to form both tri-grams and skipgrams of size three using a vector of both 0 and 1 skips.

```
text <- "an example of preprocessing techniques" tokens(text, ngrams = 3, skip = 0:1)

tokens from 1 document. text1 :
[1] "an_example_of"      "an_example_preprocessing"
[3] "an_of_preprocessing"          "an_of_techniques"
[5] "example_of_preprocessing" "example_of_techniques"
[7] "example_preprocessing_techniques"            "of_preprocessing_techniques"
```

The advantage of this approach is that these n-grams can be used in the same way as unigrams: we can make a DTM with n-grams, and perform all the types of analyses discussed above. Functions for creating a DTM in R from raw text therefore often allow the use of n-grams other than unigrams. For some analysis this can improve performance. Consider, for instance, the importance of negations and amplifiers in sentiment analysis, such as “not good” and “very bad” (Aue & Gamon, 2005). An important disadvantage, however, is that using n-grams is more computationally demanding, since there are many more unique sequences of words than individual words. This also means that more data is required to get a good estimate of the distribution of n-grams.

Another approach is to preserve the word positions after tokenization. This has three main advantages. First, the order and distance of tokens can be taken into account in analyses, enabling analyses such as the co-occurrence of words within a word window. Second, the data can be transformed into both a fulltext corpus (by pasting together the tokens) and a DTM (by dropping the token positions). This also enables the results of some text analysis techniques to be visualized in the text, such as coloring words based on a word scale model (Slapin & Proksch, 2008) or to produce browsers for topic models (Gardner et al., 2010). Third, each token can be annotated with token specific information, such as obtained from advanced NLP techniques. This enables, for instance, the use of dependency parsing to perform an analysis at the level of syntactic clauses (Van Atteveldt et al., 2017). The main disadvantage of preserving positions is that it is very memory inefficient, especially if all tokens are kept and annotations are added.

A common way to represent tokens with positions maintained is a data frame in which rows represent tokens, ordered by their position, and columns represent different variables pertaining to the token, such as the literal text, its lemma form and its POS tag. An example of this type of representation was shown above in the advanced NLP section, in the spacyr token output. Several R packages provide dedicated classes for tokens in this format. One is the koRpus package (Michalke, 2017), which specializes in various types of text statistics, in particular lexical diversity and readability. Another is corpustools (Welbers & Van Atteveldt, 2016), which focuses on managing and querying annotated tokens, and on reconstructing texts to visualize quantitative text analysis results in the original text content for qualitative investigation. A third option is tidytext (Silge & Robinson, 2016), which does not focus on this format of annotated tokens, but provides a framework for working with tokenized text in data frames.

For a brief demonstration of utilizing word positions, we perform a dictionary search with the corpustools package, that supports searching for words within a given word distance. The results are then viewed in key word in context (KWIC) listings. In the example, we look for the queries “freedom” and “americ*” within a distance of five words, using the State of the Union speeches from George W. Bush and Barack Obama.

```
install.packages("corpustools") library(corpustools)

tc <- create_tcorpus(sotu_texts, doc_column = "id")

hits <- tc$search_features('"freedom americ*"~5')

## created index for "token" column kwic <- tc$kwic(hits, ntokens = 3)

head(kwic$kwic, 3)
```

```
[1] "...making progress toward <freedom> will find <America> is their friend..."

[2] "...friends, and <freedom> in Iraq will make <America> safer for generations..."

[3] "...men who despise <freedom>, despise <America>, and aim..."
```

**Conclusion**

R is a powerful platform for computational text analysis, that can be a valuable tool for communication research. First, its well developed packages provide easy access to cutting edge text analysis techniques. As shown here, not only are most common text analysis techniques implemented, but in most cases, multiple packages offer users choice when selecting tools to implement them. Many of these packages have been developed by and for scholars, and provide established procedures for data preparation and analysis. Second, R's open source nature and excellent system for handling packages make it a convenient platform for bridging the gap between research and tool development, which is paramount to establishing a strong computational methods paradigm in communication research. New algorithms do not have to be confined to abstract and complex explanations in journal articles aimed at methodology experts, or made available through arcane code that many interested parties would not know what to do with. As an R package, algorithms can be made readily available in a standardized and familiar format.

For new users, however, choosing from the wide range of text analysis packages in R can also be daunting. With various alternatives for most techniques, it can be difficult to determine which packages are worth investing the effort to learn. The primary goal of this teacher's corner, therefore, has been to provide a starting point for scholars looking for ways to incorporate computational text analysis in their research. Our selection of packages is based on our experience as both users and developers of text analysis packages in R, and should cover the most common use cases. In particular, we advise users to become familiar with at least one established and well-maintained package that handles data preparation and management, such as quanteda, tidytext or tm. From here, it is often a small step to convert data to formats that are compatible with most of the available text analysis packages.

It should be emphasized that the selection of packages presented in this teacher's corner is not exhaustive, and does not represent which packages are the most suitable for the associated functionalities. Often, the best package for the job depends largely on specific features and problem specific priorities such as speed, memory efficiency and accuracy. Furthermore, when it comes to establishing a productive workflow, the importance of personal preference and experience should not be underestimated. A good example is the workflow of the tidytext package (Silge & Robinson, 2016), which could be preferred by people that are familiar with the tidyverse philosophy (Wickham et al., 2014). Accordingly, the packages recommended in this teacher's corner provide a good starting point, and for many users could be all they need, but there are many other great package out there. For a more complete list of packages, a good starting point is the CRAN Task View for Natural Language Processing (Wild, 2017).

Marshall McLuhan famously stated that "we shape our tools, and thereafter our tools shape us", and in science the same can be said for how tools shape our findings. We thus argue that the establishment of a strong computational methods paradigm in communication research goes hand-in-hand with embracing open-source tool development as an inherent part of scientific practice. As such, we conclude with a call for researchers to cite R packages similar to how one would cite other scientific work.[10] This gives due credit to developers, and thereby provides a just incentive for developers to publish and maintain new code, including proper testing and documentation to facilitate the correct use of code by others. Citing packages is also paramount for the transparency of research, which is especially important when using new computational techniques, where results might vary depending on implementation choices and where the absence of bugs is often not guaranteed. Just as our theories are shaped through collaboration, transparency and peer feedback, so should we shape our tools.

**Footnotes**

[1] The term "data science" is a popular buzzword related to "data-driven research" and "big data" (Provost & Fawcett, 2013).

[2] Other programming environments have similar archives, such as pip for python. However, CRAN excels in how it is strictly maintained, with elaborate checks that packages need to pass before they will be accepted.

[3] The London School of Economics and Political Science recently hosted a workshop

(http://textworkshop17.ropensci.org/), forming the beginnings of an rOpenSci special interest group for text analysis.

[4] For example, the tif (Text Interchange Formats) package (rOpenSci Text Workshop, 2017) describes and validates standards for common text data formats.

[5] https://github.com/kasperwelbers/text_analysis_in_R.

[6] For a list that includes more packages, and that is also maintained over time, a good source is the CRAN Task View for Natural Language Processing (Wild, 2017). CRAN Task Views are expert curated and maintained lists of R packages on the Comprehensive R Archive Network, and are available for various major methodological topics.

[7] http://www.tidyverse.org/.

[8] Notably, there are techniques for automatically expanding a dictionary based on the semantic space of a text corpus (Watanabe, 2017). This can be said to add an inductive layer to the approach, because the coding rules (i.e., the dictionary) are to some extent learned from the data.

[9] The term n-grams can be used more broadly to refer to sequences, and is also often used for sequences of individual characters. In this teacher's corner we strictly use n-grams to refer to sequences of words.

[10] To view how to cite a package, the citation function can be used—e.g., citation("quanteda") for citing quanteda, or citation() for citing the R project. This either provides the citation details provided by the package developer or auto-generated details.

**References**

Arnold, T. (2017a). cleannlp: A tidy data model for natural language processing [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=cleanNLP (R package version 1.9.0)

Arnold, T. (2017b). kerasR: R interface to the Keras deep learning library [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=kerasR (R package version 0.6.1)

Arnold, T., & Tilton, L. (2016). coreNLP: Wrappers around Stanford CoreNLP tools [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=coreNLP (R package version 0.4-2)

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (ranlp)* (Vol. 1, pp. 2–1).

Bates, D., & Maechler, M. (2015). Matrix: Sparse and dense matrix classes and methods [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=Matrix (R package version 1.2-3)

Benoit, K., & Matsuo, A. (2017). spacyr: R wrapper to the spacy nlp library [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=spacyr (R package version 0.9.0)

Benoit, K., & Obeng, A. (2017). readtext: Import and handling for plain and formatted text files [Computer software manual]. Retrieved from https://CRAN.R-project.org/package= readtext

Benoit, K., Watanabe, K., Nulty, P., Obeng, A., Wang, H., Lauderdale, B., & Lowe, W. (2017).

quanteda: Quantitative analysis of textual data [Computer software manual]. Retrieved from http://quanteda.io (R package version 0.99)

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Bouchet-Valat, M. (2014). Snowballc: Snowball stemmers based on the c libstemmer utf-8 library [Computer software manual]. Retrieved from https://CRAN.R-project.org/package= SnowballC (R package version 0.5.1)

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23.

Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, *173*(3), 781–800.

De Smedt, T., & Daelemans, W. (2012). "vreselijk mooi!" (terribly beautiful): A subjectivity lexicon for dutch adjectives. In *Lrec* (pp. 3568–3572).

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Fox, J., & Leanage, A. (2016). R and the journal of statistical software. *Journal of Statistical Software*, *73*(2), 1–13.

Gagolewski, M. (2017). R package stringi: Character string processing facilities [Computer software manual]. Retrieved from http://www.gagolewski.com/software/stringi/

Gardner, M. J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., & Seppi, K. (2010). The topic browser: An interactive tool for browsing topic models. In *Nips workshop on challenges of data visualization* (Vol. 2).

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. doi: 10.1093/pan/ mps028

Grun, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. doi: 10.18637/jss.v040.i13

Günther, E., & Quandt, T. (2016). Word counts and topic models: Automated text analysis methods for digital journalism research. *Digital Journalism*, *4*(1), 75–88.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & Van Atteveldt, W. (2014). Rtexttools: Automatic text classification via supervised learning [Computer software manual].

Retrieved from https://CRAN.R-project.org/package=RTextTools (R package version

1.4.2)

Lang, D. T., & the CRAN Team. (2017). Xml: Tools for parsing and generating xml within r and s-plus [Computer software manual]. Retrieved from https://CRAN.R-project.org/ package=XML

Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, *46*(1), 423–444.

Manning, C. D., Manning, C. D., Raghavan, P., Raghavan, P., Schütze, H., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60).

McCarthy, P. M., & Jarvis, S. (2010). Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, *42*(2), 381–392.

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in r. *Journal of statistical software*, *25*(5), 1–54.

Michalke, M. (2017). korpus: An r package for text analysis [Computer software manual]. Retrieved from https://reaktanz.de/?c=hacking&s=koRpus ((Version 0.10-2))

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expertat Systems with Applications*, *40*(10), 4241 - 4251. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417413000328 doi: http:// dx.doi.org/10.1016/j.eswa.2013.01.019

Mullen, L. (2016a). textreuse: Detect text reuse and document similarity [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=textreuse (R package version 0.1.4)

Mullen, L. (2016b). tokenizers: A consistent interface to tokenize natural language text [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tokenizers (R package version 0.1.4)

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects [Computer software manual]. Retrieved from https://arxiv.org/abs/1403.2805

Ooms, J. (2017a). antiword: Extract text from microsoft word documents [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=antiword

Ooms, J. (2017b). pdftools: Text extraction, rendering and converting of pdf documents [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=pdftools Porter, M. F. (2001). *Snowball: A language for stemming algorithms.*

Proksch, S.-O., & Slapin, J. B. (2009). How to avoid pitfalls in statistical analysis of political texts: The case of germany. *German Politics*, *18*(3), 323–344.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, *1*(1), 51–59.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand,D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.

rOpenSci Text Workshop. (2017). tif: Text interchange format [Computer software manual]. Retrieved from https://github.com/ropensci/tif

Schuck, A. R., Xezonakis, G., Elenbaas, M., Banducci, S. A., & De Vreese, C. H. (2011). Party contestation and europe on the news agenda: The 2009 european parliamentary elections. *Electoral Studies*, *30*(1), 41–52.

Schultz, F., Kleinnijenhuis, J., Oegema, D., Utz, S., & Van Atteveldt, W. (2012). Strategic framing in the bp crisis: A semantic network analysis of associative frames. *Public Relations Review*, *38*(1), 97–107.

Selivanov, D. (2016). text2vec: Modern text mining framework for r [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=text2vec (R package version 0.4.0)

Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, *1*(3). Retrieved from http://dx.doi.org/10.21105/joss.00037 doi: 10.21105/ joss.00037

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705–722.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267–307.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 2454. doi: 10.1177/0261927X09351676

TIOBE. (2017). *The R Programming Language.* Retrieved from https://www.tiobe.com/tiobe

-index/r/

Van Atteveldt, W. (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media content (dissertation)*. BookSurge.

Van Atteveldt, W., Sheafer, T., Shenhav, S. R., & Fogel-Dror, Y. (2017). Clause analysis: using syntactic information to automatically extract source, subject, and predicate from texts with an application to the 2008–2009 gaza war. *Political Analysis*, 1–16.

Vliegenthart, R., Boomgaarden, H. G., & Van Spanje, J. (2012). Anti-immigrant party support and media visibility: A cross-party, over-time perspective. *Journal of Elections, Public Opinion & Parties*, *22*(3), 315–358.

Watanabe, K. (2017). The spread of the kremlin's narratives by a western news agency during the ukraine crisis. *The Journal of International Communication*, *23*(1), 138–158.

Welbers, K., & Van Atteveldt, W. (2016). Corpustools [Computer software manual]. Retrieved from http://github.com/kasperwelbers/corpustools (R package version 0.201)

Welbers, K., Van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2016). A gatekeeper among gatekeepers: News agency influence in print and online newspapers in the netherlands. *Journalism Studies*, 1–19.

Wickham, H. (2017). stringr: Simple, consistent wrappers for common string operations [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., & Bryan, J. (2017). readxl: Read excel files [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=readxl

Wickham, H., et al. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23.

Wild, F. (2017). Cran task view: Natural language processing. *CRAN*. Version: 2017-01-17, https://CRAN.R-project.org/view=NaturalLanguageProcessing.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, pp. 412–420).

Terminológiai gyűjtemény

| sorszám | Forrásnyelvi terminus | Célnyelvi terminus 1 | Célnyelvi terminus 2 (szinoníma) |
|---|---|---|---|
| 1 | advanced | haladó | összetett |
| 2 | analysis | elemzés | |
| 3 | classifier | osztályozó | |
| 4 | cleaning | tisztítás | |
| 5 | corpus | korpusz | |
| 6 | dependency | kapcsolati | |
| 7 | dictionary | szótár | |
| 8 | figure | ábra | |
| 9 | file | fájl | |
| 10 | filtering | szűrés | |
| 11 | funcition | függvény | funkció |
| 12 | machine learning | gépi tanulás | |
| 13 | modelling | modellezés | |
| 14 | NLP | természetes nyelvfeldolgozás | |
| 15 | object | objektum | fájl |
| 16 | package | program | |
| 17 | statistics | statisztika | statisztikák |
| 18 | supervised | felügyelt | |
| 19 | syntax | szintaktika | |
| 20 | Teacher’s corner | Módszertani Segédlet | |
| 21 | term | teminus | kifejezés |
| 22 | text | szöveg | |
| 23 | topic | téma | |

A fordítandó szöveg elejének fordítása a memoQ program alkalmazásával

| **kepesito_forditas01.docx**<br>**CAUTION: Do not change segment ID or source text**<br>**V9.12.9 MQ990911 1afa04d8-6215-4086-8b1e-cbd7ca77b663** | | | | |
|---|---|---|---|---|
| **ID** | **English (United Kingdom)** | **Hungarian** | **Comment** | **Status** |
| 1 | [1}Abstract. | **[1}Összefoglalás** | | Edited |
| 2 | Computational text analysis has become an exciting research field with many applications in communication research. | A számítógépes szövegelemzés izgalmas kutatási területté vált, melynek számos alkalmazása van a kommunikáció kutatásában. | | Confirmed |
| 3 | It can be a difficult method to apply, however, because it requires knowledge of various techniques, and the software required to perform most of these techniques is not readily available in common statistical software packages. | Felhasználása bonyolult lehet, mert különböző technikák ismeretét teszi szükségessé, és a legtöbb módszer alkalmazásához szükséges szoftverek nem érhetőek el közvetlenül az általánosan használt statisztikai szoftver programokból. | | Confirmed |
| 4 | In this teacher’s corner, we address these barriers by providing an overview of general steps and operations in a computational text analysis project, and demonstrate how each step can be performed using the R statistical software. | Ebben a módszertani segédletben ezeket az akadályokat tárgyaljuk és átfogó áttekintést adunk egy számítógépes szövegelemzési projekt általános kérdéseiről és műveleteiről, bemutatva, hogy minden egyes lépés hogyan hajtható végre az R statisztikai szoftver felhasználásával. | A "Teacher's corner" a magyar olvasó számára értelmezhetetlen. Ez nem tanároknak szól és nem szakkör. Ehelyett azt írtam, ami a valódi műfaja: módszertani segédlet. | Confirmed |
| 5 | As a popular open-source platform, R has an extensive user community that | Népszerű, nyílt forrású platformként az R kiterjedt felhasználói közösséggel | | Confirmed |

| | | | | |
|---|---|---|---|---|
| | develops and maintains a wide range of text analysis packages. | rendelkezik, mely a szövegelemző programok széles spektrumát fejleszti és tartja fenn. | | |
| 6 | We show that these packages make it easy to perform advanced text analytics. | Bemutatjuk, hogy ezen programok hogyan könnyítik meg a bonyolultabb szövegelemzések elvégzését. | | Confirmed |
| 7 | With the increasing importance of computational text analysis in communication research (Boumans & Trilling, 2016; Grimmer & Stewart, 2013), many researchers face the challenge of learning how to use advanced software that enables this type of analysis. | A számítógépes szövegelemzés növekvő jelentőségével párhuzamosan a kommunikáció kutatásban számos kutató szembesül azzal a kihívással (Boumans & Trilling, 2016; Grimmer & Stewart, 2013), hogy meg kell tanulnia alkalmazni az ilyen típusú elemzés elvégzésére alkalmas, fejlett szoftvereket. | | Confirmed |
| 7 | Currently, one of the most popular environments for computational methods and the emerging field of “data science” {1][2}1 {3][4}is the R statistical software (R Core Team, 2017). | Jelenleg a számítási módszerek és a feltörekvő 1][2}„adattudomány” {1][2} egyik legnépszerűbb környezete az R statisztikai szoftver (R Core Team, 2017). | | Edited |
| 8 | However, for researchers that are not well-versed in programming, learning how to use R can be a challenge, and performing text analysis in particular can seem daunting. | A programozásban kevésbé járatos kutatók számára azonban kihívást jelenthet az R használata, és a szövegelemzés megvalósítása különösen ijesztőnek tűnhet. | | Confirmed |
| 9 | In this teacher’s corner, we show that performing text analysis in R is not as hard as some might fear. | Ebben a módszertani segédletben bemutatjuk, hogy a szövegelemzés végrehajtása nem olyan nehéz, mint amennyire néhányan tartanak tőle. | | Confirmed |

| | | | | |
|---|---|---|---|---|
| 10 | We provide a step-bystep introduction into the use of common techniques, with the aim of helping researchers get acquainted with computational text analysis in general, as well as getting a start at performing advanced text analysis studies in R. | Lépésről lépésre mutatjuk be a leggyakrabban alkalmazott eljárások használatát, azzal a céllal, hogy segítsük a kutatókat a számítógépes szövegelemzés általános megismerésében és az R-ben végzendő fejlett szövegelemzési munkák megvalósításában. | | Confirmed |
| 11 | R is a free, open-source, cross-platform programming environment. | Az R szabadon felhasználható, nyílt forráskódú, platformfüggetlen programozási környezet. | | Confirmed |
| 12 | In contrast to most programming languages, R was specifically designed for statistical analysis, which makes it highly suitable for data science applications. | Más programnyelvekkel ellentétben az R-t kifejezetten statisztikai elemzésre tervezték, ezért kiválóan alkalmas adattudományi alkalmazásokra. | | Confirmed |
| 13 | Although the learning curve for programming with R can be steep, especially for people without prior programming experience, the tools now available for carrying out text analysis in R make it easy to perform powerful, cutting-edge text analytics using only a few simple commands. | Bár az R programozásának tanulási görbéje – különösen azok számára, akik nem rendelkeznek programozási tapasztalatokkal – meredek lehet, az R-ben történő szövegelemzés megvalósítását szolgáló, jelenleg rendelkezésre álló eszközök megkönnyítik a mindössze néhány egyszerű parancs alkalmazásával végrehajtott hatékony, korszerű szövegelemzést. | | Confirmed |
| 14 | One of the keys to R's explosive growth (Fox & Leanage, 2016; TIOBE, 2017) has been its densely populated collection of extension software libraries, known in | Az R robbanásszerű bővülésének (Fox & Leanage, 2016; TIOBE, 2017) egyik kulcsa a nagyszámú kiegészítő szoftverkönyvtár, az úgynevezett | Amint arra a dolgozat egy másik pontján utalok, a package szó szó szerinti fordítását zavarosnak tartom, ehelyett alkalmazom egységesen a | Confirmed |

| | | | | |
|---|---|---|---|---|
| | R terminology as *{1][2}packages*{3][4}, supplied and maintained by R's extensive user community. | programok, melyeket az R kiterjedt felhasználói közössége készít és tart fenn. | program szót, hiszen ez a dokumentum is csak „programozási környezetnek" tekinti az R-t. | |
| 15 | Each package extends the functionality of the base R language and core packages, and in addition to functions and data must include documentation and examples, often in the form of vignettes demonstrating the use of the package. | Minden egyes program tovább bővíti az R-nyelv és az alapprogramok felhasználási lehetőségeit. Az új funkciók és adatok hozzáadásához dokumentációra és példákra is szükség van, melyek gyakran címkék formájában mutatják be a program alkalmazását. | | Confirmed |
| 16 | The best-known package repository, the Comprehensive R Archive Network (CRAN), currently has over 10,000 packages that are published, and which have gone through an extensive screening for procedural conformity and cross-platform compatibility before being accepted by the archive.{1][2}2 {3][4} | A legismertebb program-könyvtár az átfogó R-archívum [Comprehensive R Archive Network (CRAN)], mely jelenleg több mint 10 000 közzétett, az eljárások összhangja és a platformfüggetlenség szempontjából előzetesen és széleskörűen ellenőrzött programot tartalmaz. | | Confirmed |
| 17 | R thus features a wide range of inter-compatible packages, maintained and continuously updated by scholars, practitioners, and projects such as RStudio and rOpenSci. | Így az R kiterjedt, egymással komptatibilis programokból áll, melyeket a tudósok, a gyakorló szakemberek és az olyan projektek, mint például az Rstudio és az rOpenSci folyamatosan karbantartanak és frissítenek. | | Confirmed |
| 18 | Furthermore, these packages may be installed easily and safely from within the R environment using a single command. | Fontos tudnunk, hogy ezeket a programokat könnyen és biztonságosan, egy parancs | | Confirmed |

| | | | | |
|---|---|---|---|---|
| | | alkalmazásával lehet installálni az R-környezetből. | | |
| 19 | R thus provides a solid bridge for developers and users of new analysis tools to meet, making it a very suitable programming environment for scientific collaboration. | Ebből adódóan az R szilárd, a fejlesztők és az új elemzési eszközök felhasználói közötti találkozási felületet képez, ezért ez a programozási környezet rendkívül alkalmas a tudományos együttműködésre. | | Confirmed |
| 20 | Text analysis in particular has become well established in R. | A szövegelemzés különösen jól illeszkedik az R-be. | | Confirmed |
| 21 | There is a vast collection of dedicated text processing and text analysis packages, from low-level string operations (Gagolewski, 2017) to advanced text modeling techniques such as fitting Latent Dirichlet Allocation models (Blei, Ng, & Jordan, 2003; Roberts et al., 2014)—nearly 50 packages in total at our last count. | A legutóbbi felmérésünk alapján összesen közel ötven programot magába foglaló, kifejezetten a szövegfeldolgozást és a szövegelemzést szolgáló hatalmas gyűjteménye van, az alacsony szintű karakterlánc-kezelési műveletektől (Gagolewski, 2017) az olyan magas szintű szöveg-modellezési eljárásokig , mint például a látens Dirichlet allokációs modellek (Blei, Ng, & Jordan, 2003; Roberts et al., 2014). | | Confirmed |
| 22 | Furthermore, there is an increasing effort among developers to cooperate and coordinate, such as the rOpenSci special interest group.{1][2}3 {3][4} | A fejlesztők között fokozódó erőfeszítések tapasztalhatók az együttműködésre és az erőfeszítések összehangolására, ilyen például az rOpenSci szakosított együttműködés. | | Confirmed |
| 23 | One of the main advantages of performing text analysis in R is that it is often possible, and relatively easy, to | Az R-ben végrehajtott szövegelemzés egyik fő előnye, hogy gyakran lehetséges és viszonylag könnyű a | | Confirmed |

| | | | | |
|---|---|---|---|---|
| | switch between different packages or to combine them. | különböző programok közötti váltás, vagy azok együttes felhasználása. | | |
| 24 | Recent efforts among the R text analysis developers' community are designed to promote this interoperability to maximize flexibility and choice among users.{1][2}4 {3][4}As a result, learning the basics for text analysis in R provides access to a wide range of advanced text analysis features. | A R szövegelemzésfejlesztő közösségének legutóbbi erőfeszítési arra irányulnak, hogy elősegítsék az egyes programok közötti interoperabilitást a rugalmasság és a felhasználó szabadságának maximalizálása érdekében. Így, ha elsajátítjuk az R-ben végzett szövegelemzés alapjait, a fejlettebb szövegelemzési lehetőségek egész sora nyílik meg előttünk. | | Confirmed |
| 25 | **{1][2}Structure of this Teacher's Corner** | **A Módszertani Segédlet felépítése** | | Confirmed |
| 26 | {1][2}This teacher's corner covers the most common steps for performing text analysis in R, from data preparation to analysis, and provides easy to replicate example code to perform each step. | Ez a módszertani segédlet az R-ben végzett leggyakoribb szövegelemzési lépéseket tartalmazza, az adat-előkészítéstől az elemzésig, és minden egyes lépéshez könnyen megismételhető példa-kódokat nyújt. | | Confirmed |
| 27 | The example code is also digitally available in our online appendix, which is updated over time.{1][2}5 {3][4} | A példa-kódok digitálisan is elérhetők a folyamatosan frissített online függelékben is. | | Confirmed |
| 28 | We focus primarily on bag-of-words text analysis approaches, meaning that only the frequencies of words per text are used and word positions are ignored. | Elsősorban a szózsák típusú szövegelemzési megközelítésre összpontosítunk. Ez azt jelenti, hogy csak a szövegenként előforduló szavak gyakoriságát használjuk fel és nem vesszük figyelembe azok helyzetét. | | Confirmed |

| | | | | |
|---|---|---|---|---|
| | Although this drastically simplifies text content, research and many real-world applications show that word frequencies alone contain sufficient information for many types of analysis (Grimmer & Stewart, 2013). | Ez erőteljesen egyszerűsíti ugyan a szöveg tartalmát, de a kutatások és számos gyakorlati alkalmazás azt mutatja, hogy a szavak gyakorisága önmagában elégséges információt tartalmaz sokféle elemzés elvégzéséhez (Grimmer & Stewart, 2013). | | Confirmed |
| | Table 1 presents an overview of the text analysis operations that we address, categorized in three sections. | Az 1. táblázatban áttekintés adunk a bemutatott szövegelemzési műveletekről, három csoportra osztva azokat. | | Confirmed |
| | In the *{1][2}data preparation* {3][4}section we discuss five steps to prepare texts for analysis. | Az *{3]{1]adatelőkészítésről szóló* [2}[4} fejezetben az elemzésre történő szövegelőkészítés öt lépését tárgyaljuk. | | Edited |
| | The first step, *{1][2}importing text*{3][4}, covers the functions for reading texts from various types of file formats (e.g., txt, csv, pdf) into a *{1][2}raw text corpus* {3][4}in R. | Az első lépés a{1][4} *szöveg importálás*{3][2}a, mely a szöveg különböző file-formátumokból (pl. txt, csv, pdf) a {1][2} *nyers szöveg-testbe (korpuszba)* {történő beolvasásához szükséges feladatokat foglalja magába1][2} az R-be. | | Edited |
| | The steps *{1][2}string operations* {3][4}and *{1][2}preprocessing* {3][4}cover techniques for manipulating raw texts and processing them into *{1][2}tokens* {3][4}(i.e., units of text, such as words or word stems). | Ezek a *{1][4}karakterlánc-műveletek*{1][4} és az *{1][4}előzetes feldolgozás* {1][4}azokat a nyers szövegkezelési és előfeldolgozási technikákat jelentik, melyek célja a nyers szöveg átalakítása és feldolgozása *{1][4}tokenekké (szótövekké)*{1][4} (azaz | | Edited |

| | | | | |
|---|---|---|---|---|
| | | szövegelemekké, például szavakká és szótövekké). | | |
| | The tokens are then used for creating the *{1][2}document-term matrix* {3][4}(DTM), which is a common format for representing a bag-of-words type corpus, that is used by many R text analysis packages. | Ezt követően a példányokat a *{1][2}dokumentum-kifejezés mátrix*{1][2} (document-term matrix, DTM) létrehozására használjuk fel. Ez általánosan használt formája a szózsák-típusú szövegtörzsek megjelenítésének, melyeket számos R szövegelemző program használ. | | Edited |
| | Other nonbag- of-words formats, such as the tokenlist, are briefly touched upon in the *{1][2}advanced topics* {3][4}section. | Egyéb, nem szózsák alapú formátumokat, például a példánylistákat ugyancsak bemutatunk a *{3][2}"bonyolultabb kérdések"*{1][2} fejezetben. | | Edited |
| | Finally, it is a common step to *{1][2}filter and weight* {3][4}the terms in the DTM. | Végül gyakori az egyes kifejezések *{1][2}szűrése és súlyozása*{1][2} is a DTM-ben (document-term matrix). | | Edited |
| | These steps are generally performed in the presented sequential order (see Figure 1 for conceptual illustration). | Ezeket a lépéseket általában a bemutatott sorrendben hajtjuk végre (a fogalmi illusztrációhoz lásd az 1. ábrát). | | Confirmed |
| | As we will show, there are R packages that provide convenient functions that manage multiple data preparation steps in a single line of code. | Amint látni fogjuk, vannak R-programok, melyek megfelelő funkciókkal rendelkeznek ahhoz, hogy többféle adatelőkészítő lépés egyetlen programsorral megvalósítható legyen. | | Confirmed |
| | Still, we first discuss and demonstrate each step separately to provide a basic understanding of the purpose of each | Ennek ellenére mi először minden egyes lépést külön-külön tárgyalunk és mutatunk be, annak érdekében, hogy | | Confirmed |

| | | | | |
|---|---|---|---|---|
| | step, the choices that can be made and the pitfalls to watch out for. | alapszinten megértsük az egyes lépések célját, a választási lehetőségeket és felhívjuk a figyelmet a buktatókra. | | |
| | The *{1][2}analysis* {3][4}section discusses four text analysis methods that have become popular in communication research (Boumans & Trilling, 2016) and that can be performed with a DTM as input. | Az{3][4} *elemzés*{3][4} fejezetben a kommunikációkutatásban népszerű négy szövegelemzési módszert tárgyalunk ((Boumans & Trilling, 2016), melyeket a DTM-input felhasználásával hajthatunk végre. | | Confirmed |
| | Rather than being competing approaches, these methods have different advantages and disadvantages, so choosing the best method for a study depends largely on the research question, and sometimes different methods can be used complementarily (Grimmer & Stewart, 2013). | Ahelyett, hogy ezeket egymással versengő megközelítéseknek tekintenénk, látnunk kell, hogy az egyes módszereknek különböző előnyeik és hátrányaik vannak, és ezért a kutatás szempontjából legjobb módszer kiválasztása nagymértékben függ a kutatási kérdéstől, és néha a különböző módszerek kiegészíthetik egymást (Grimmer & Stewart, 2013). | | Confirmed |
| | Accordingly, our recommendation is to become familiar with each type of method. | Ennek megfelelően azt javasoljuk, hogy mindegyik módszerrel érdemes megismerkedni. | | Confirmed |
| | To demonstrate the general idea of each type of method, we provide code for typical analysis examples. | Annak érdekében, hogy bemutassuk, milyen általános elvre épülnek az egye módszertípusok, a tipikus elemzési példák kódját is megadjuk. | | Confirmed |

| | | | | |
|---|---|---|---|---|
| 45 | Furthermore, it is important to note that different types of analysis can also have different implications for how the data should be prepared. | Vegyük figyelembe azt is, hogy a különböző elemzések eltérő adatelőkészítést igényelhetnek. | | Confirmed |
| 46 | For each type of analysis we therefore address general considerations for data preparation. | Ebből adódóan minden egyes elemzési típus esetében az adatelőkészítéssel kapcsolatos megfontolásokat is tárgyaljuk. | | Confirmed |
| 47 | Finally, the additional *{1][2}advanced topics* {3][4}section discusses alternatives for data preparation and analysis that require external software modules or that go beyond the bag-of-words assumption, using word positions and syntactic relations. | Végül, a *{1][2}haladó témakörök*{3][4} fejezetben alternatív adat-előkészítési és elemzési módszereket tárgyalunk, melyek külső szoftver-modulok alkalmazását teszik szükségessé vagy túlmennek a szózsák-feltételezésen, mert felhasználják a szavak helyzetét és szintaktikai kapcsolatait is. | | Confirmed |
| 48 | The purpose of this section is to provide a glimpse of alternatives that are possible in R, but might be more difficult to use. | Ennek a fejezetnek az a célja, hogy betekintést nyújtson az R-rel elérhető alternatív, olykor bonyolultabban használható megoldásokra. | | Confirmed |
| 49 | Within each category we distinguish several groups of operations, and for each operation we demonstrate how they can be implemented in R. | Minden egyes kategórián belül néhány műveletcsoportot különítünk el, és minden egyes műveletre bemutatjuk, hogy az hogyan valósítható meg R-ben. | | Confirmed |
| 50 | To provide parsimonious and easy to replicate examples, we have chosen a specific selection of packages that are easy to use and broadly applicable. | Annak érdekében, hogy tömör és egyszerűen megismételhető példákat mutassunk be, olyan programokat választottunk, melyek egyszerűen | | Confirmed |

| | | | | |
|---|---|---|---|---|
| | | használhatók és széles körben alkalmazhatók. | | |
| 51 | However, there are many alternative packages in R that can perform the same or similar operations. | Tudnunk kell azonban, hogy számos alternatív program is elérhető, melyek azonos vagy hasonló műveletek végrehajtására képesek. | | Confirmed |
| 52 | Due to the open-source nature of R, different people from often different disciplines have worked on similar problems, creating some duplication in functionality across different packages. | Az R nyílt forráskód jellegéből adódóan a különböző kutatók gyakran egymásól eltérő tudományterületekről dolgoznak hasonló problémákon, és így az egyes funkciók között átfedések alakulhatnak ki az egyes programok esetén. | | Confirmed |
| 53 | This also offers a range of choice, however, providing alternatives to suit a user's needs and tastes. | Ez egyúttal a felhasználói igénynek és ízlésnek megfelelő alternatívákat kínál. | | Confirmed |
| 54 | Depending on the research project, as well as personal preference, other packages might be better suited to different readers. | A kutatási projekttől és az egyéni preferenciáktól függően lehetséges, hogy más programok jobban megfelelnek a különböző olvasóknak. | | Confirmed |
| 55 | While a fully comprehensive review and comparison of text analysis packages for R is beyond our scope here—especially given that existing and new packages are constantly being developed—we have tried to cover, or at least mention, a variety of alternative packages for each text analysis operation. | Jóllehet az R szövegelemző programok teljeskörű áttekintése és összehasonlítása meghaladja vizsgálataink körét – különösen azért, mert folyamatosan fejlesztik a meglévő és az új programokat – mégis megkíséreltük legalább az említés szintjén érinteni az egyes szövegelemzési műveletek során alkalmazható alternatív programokat. | | Confirmed |

| | | | | |
|---|---|---|---|---|
| 56 | {1][2}6 {3][4}In general, these packages often use the same standards for data formats, and thus are easy to substitute or combine with the other packages discussed in this teacher’s corner.{1] | Általában ezen programok gyakran használják ugyanazt az standard adatformátumot, és így könnyen helyettesíthetjük vagy kombinálhatjuk azokat a jelen módszertani segédletben bemutatott programokkal. | | Confirmed |

A cikk egy további részének gépi fordítása, részleges és teljes utószerkesztése, és az ahoz kapcsolódó fordítói megjegyzések

| Eredeti szöveg | Gépi fordítás | Részleges utószerkesztés | Teljes utószerkesztés | Fordítói megjegyzés |
|---|---|---|---|---|
| The advantage of this representation is that it allows the data to be analyzed with vector and matrix algebra, effectively moving from text to numbers. | Ennek a reprezentációnak az az előnye, hogy lehetővé teszi az adatok vektor- és mátrixalgebrai elemzését, gyakorlatilag a szövegről a számokra való áttérést. | Ennek a reprezentációnak az az előnye, hogy lehetővé teszi az adatok vektor- és mátrixalgebrai elemzését, gyakorlatilag a szövegről a számokra történő áttérést. nem szükséges a részleges utószerkesztés | Ennek a megjelenítésnek az az előnye, hogy lehetővé teszi az adatok statisztikai elemzését, azaz a szövegről a számokra történő áttérést. | A vector-és mátrixalgebrai elemzést nem önmagáért végezzük, hanem azért, hogy azokból statisztikai információkat gyűjtsünk. Ezért a fordításban igyekeztem sokkal tömörebb lenni, mint az eredeti szöveg. |
| Furthermore, with the use of special matrix formats for sparse matrices, text data in a DTM format is very memory efficient and can be analyzed with highly optimized operations. | Továbbá a ritka mátrixok speciális mátrixformátumainak használatával a DTM formátumú szöveges adatok nagyon memóriahatékonyak, és rendkívül optimalizált műveletekkel elemezhetők. | A ritka mátrixok speciális mátrixformátumainak használatával a DTM formátumú szöveges adatok nagyon memóriahatékonyak, és rendkívül optimalizált műveletekkel elemezhetők. | A ritka mátrixok speciális mátrixformátumainak használatával a DTM -ben tárolt szöveges adatok nagyon hatékony memóriakezelést és optimalizált műveletvégzést tesznek lehetővé. | Matematikailag és nyelvtanilag az eredeti mondat hibás. Egy rendszer működésének optimuma az a pont, ami az adott feltételek mellett a lehető legkedvezőbb. Ezért “rendkívül optimalizált” műveletekről beszélni érelemetlen. A rikta matrix a magyar nyelvű szakirodalomban elterjedt szó. Szerintem szebb lenne a gyér mártix elnezvezés, mert szemléletesebb, de ez nem terjedt el. Kihagytam, a továbbá kötőszót, valamint a memóriahatékonyak szó esetében nyelvtani |

| | | | | |
|---|---|---|---|---|
| | | | | csrével (melléknév használata), éltem. |
| Two of the most established text analysis packages in R that provide dedicated DTM classes are tm and quanteda. | Az R két legelterjedtebb szövegelemző csomagja, amelyek dedikált DTM osztályokat biztosítanak, a tm és a quanteda | Az R két legelterjedtebb szövegelemző programja, amelyek dedikált DTM fájlokat biztosítanak, a tm és a quanteda | Az tm és a quanteda az R két legelterjedtebb szövegelemző programja, amelyekkel speciális DTM fájlok hozthatók létre | Értelemzhetőbbé tettem a szórendet és jobb jelzőt igyekeztem találni a DTM elé, valamint a pacakges terminust a szakma által használt terminusra cseréltem. |
| Of the two, the venerable tm package is the more commonly used, with a user base of almost 10 years (Meyer, Hornik, & Feinerer, 2008) and several other R packages using its DTM classes (DocumentTermMatrix and TermDocumentMatrix) as inputs for their analytic functions. | A kettő közül a tiszteletre méltó tm csomag a leggyakrabban használt, közel 10 éves felhasználói bázissal (Meyer, Hornik, & Feinerer, 2008) és számos más R csomag használja a DTM osztályait (DocumentTermMatrix és TermDocumentMatrix) elemző függvényeik bemeneteként | A kettő közül a jóöreg tm csomag a leggyakrabban használt, közel 10 éves felhasználói bázissal (Meyer, Hornik, & Feinerer, 2008), továbbá számos más R csomag is használja a programmal létrheozott DTM fájlokat (DocumentTermMatrix és TermDocumentMatrix) elemző funkcióik bemeneteként | A kettő közül a jóöreg tm program a leggyakrabban alkalmazott, közel 10 éves felhasználói tapasztalattal (Meyer, Hornik, & Feinerer, 2008). Az általa létrehozott DTM fájlokat (DocumentTermMatrix és TermDocumentMatrix) számos más R program is használja az elemző funkcióik bemeneteként | Az utószerkesztésben a rendkívül hosszú mondatot két egyszerűbbre bontottam. |
| The quanteda package is a more recently developed package, built by a team supported by an ERC grant to provide state-of-the-art, high performance text analysis. | A quanteda csomag egy újabb fejlesztésű csomag, amelyet egy ERC-ösztöndíjjal támogatott csapat hozott létre, hogy a legkorszerűbb, nagy teljesítményű szövegelemzést nyújtsa | A quanteda program újabb fejlesztésű, amelyet egy ERC-ösztöndíjjal támogatott csapat hozott létre, azért, hogy a legkorszerűbb, nagy teljesítményű szövegelemzést nyújtsa | A quanteda program újabb fejlesztésű, amelyet egy ERC-ösztöndíjjal támogatott csapat azért hozott létre, hogy a legkorszerűbb, nagy teljesítményű szövegelemzést valósítsa meg | Módosítottam a szórenden. Lexikai cserét végeztem. |
| Its sparse DTM class, known as a dfm or document-feature matrix, is based on the powerful Matrix package | Gyér DTM-osztálya, amelyet dfm vagy dokumentum-tulajdonságmátrix néven ismerünk, háttértárként a | Az általa létrehozott, a ritka mátrixok kezelését lehetővé tevő DTM fájl, amelyet dfm vagy dokumentum-tulajdonságmátrix néven | Az programmal létrehozott, a ritka mátrixok kezelését lehetővé tevő DTM fájl, amelyet dfm vagy dokumentum- | A gépi fordítást ki kellett egészíteni azért, hogy könnyebben értelmezhető legyen. |

| | | | | |
|---|---|---|---|---|
| (Bates & Maechler, 2015) as a backend, but includes functions to convert to nearly every other sparse document-term matrix used in other R packages (including the tm formats). | nagy teljesítményű Matrix csomagra (Bates & Maechler, 2015) épül, de tartalmaz olyan függvényeket, amelyekkel szinte minden más, más R-csomagokban használt gyér dokumentum-term mátrixra (beleértve a tm formátumokat is) konvertálható | ismerünk, háttértárként a nagy teljesítményű Matrix programra (Bates & Maechler, 2015) épül, de tartalmaz olyan funkciókat, amelyekkel szinte minden más, más R-programban használt ritka dokumentum-term mátrix (beleértve a tm formátumokat is) is konvertálható | tulajdonságmátrix néven ismerünk, háttértárként a nagy teljesítményű Matrix programra (Bates & Maechler, 2015) épül, de tartalmaz olyan funkciókat is, amelyekkel szinte minden más, R-programban használt ritka dokumentum-term mátrix (beleértve a tm formátumokat is) is konvertálható | |
| The performance and flexibility of quanteda's dfm format lends us to recommend it over the tm equivalent. | A quanteda dfm formátumának teljesítménye és rugalmassága miatt ajánljuk a tm megfelelőjével szemben | A quanteda dfm fájljának teljesítménye és rugalmassága miatt ennek használatát ajánljuk a tm megfelelő funkciójával szemben | A quanteda dfm fájlját teljesítménye és rugalmassága miatt ennek használatát ajánljuk a tm megfelelő funkciójával szemben | Lexikai cserét és beszúrást alkalmaztam a könnyebb érthetőség miatt |
| Another notable alternative is the tidytext package (Silge & Robinson, 2016). | Egy másik figyelemre méltó alternatíva a tidytext csomag (Silge & Robinson, 2016). | Egy másik figyelemre méltó alternatíva a tidytext program (Silge & Robinson, 2016). | Egy további figyelemre méltó alternatíva a tidytext program (Silge & Robinson, 2016). | A gépi fordítás majdnem hibátlan volt. |
| This is a text analysis package that is part of the Tidyverse7—a collection of R packages with a common philosophy and format. | Ez egy szövegelemző csomag, amely része a Tidyverse7 - egy közös filozófiájú és formátumú R-csomagok gyűjteményének. | Ez egy olyan szövegelemző program, amely része a Tidyversenek, azaz egy közös filozófiájú és formátumú R-programok gyűjteményének. | Ez egy olyan szövegelemző program, amely része a Tidyversenek, azaz egy közös elven és formátumon alapuló R-programok gyűjteményének. | Lexikai cserét végeztem |
| Central to the Tidyverse philosophy is that all data is arranged as a table, where (1) "each variable forms a | A Tidyverse filozófiájának központi eleme, hogy minden adat táblázatba rendeződik, ahol (1) "minden változó egy | A Tidyverse filozófiájának központi eleme, hogy minden adat táblázatba rendeződik, ahol (1) "minden változó egy | A Tidyverse filozófiájának központi eleme, hogy minden adat táblázatba rendeződik, ahol (1) "minden változó egy | A gépi fordítás elfogadható |

| | | | | |
|---|---|---|---|---|
| column", (2) "each observation forms a row", and (3) "each type of observational unit forms a table" (Wickham et al., 2014, 4). | oszlopot", (2) "minden megfigyelés egy sort" és (3) "minden megfigyelési egységtípus egy táblázatot" alkot (Wickham et al. , 2014, 4). | oszlopot", (2) "minden megfigyelés egy sort" és (3) "minden adatgyűjtés-sorozat egy táblázatot" alkot (Wickham et al., 2014, 4). | oszlopot", (2) "minden megfigyelési egység egy sort" és (3) "minden adatgyűjtés-sorozat egy táblázatot" alkot (Wickham et al. , 2014, 4). | |
| As such, tidytext does not strictly use a document term matrix, but instead represents the same data in a long format, where each (non-zero) value of the DTM is a row with the columns document, term, and count (note that this is essentially a triplet format for sparse matrices, with the columns specifying the row, column and value). | Mint ilyen, a tidytext nem szigorúan dokumentum-terminus mátrixot használ, hanem ugyanezeket az adatokat egy hosszú formátumban ábrázolja, ahol a DTM minden (nem nulla) értéke egy sor, amelynek oszlopai a dokumentum, a terminus és a szám (megjegyzendő, hogy ez lényegében a ritka mátrixok triplett formátuma, ahol az oszlopok megadják a sort, az oszlopot és az értéket). | Mint ilyen, a tidytext nem szigorúan dokumentum-terminus mátrixot használ, hanem ugyanezeket az adatokat egy olyan hosszú formátumban tárolja, ahol a DTM minden (nem nulla) értéke egy sor, és melynek oszlopai a dokumentum, a terminus és a gyakoriság (megjegyzendő, hogy ez lényegében a ritka mátrixok triplett formátuma, ahol az oszlopokba rendezett információk adják meg a sort, az oszlopot és a gyakoriságot). | Mint ilyen, a tidytext nem szigorúan dokumentum-term mátrixot használ, hanem ugyanezeket az adatokat egy olyan hosszú formátumban tárolja, ahol a DTM minden (nem nulla) értéke egy sor, és melynek oszlopai a dokumentum, a terminus és a gyakoriság (megjegyzendő, hogy ez lényegében a ritka mátrixok triplett formátuma, ahol az oszlopokba renezett információk adják meg a sort, az oszlopot és az értéket). | A gépi fordítás majdnem teljesen megfelel, de a kiegészítések értelmezhetőbbé teszik a szöveget. Lexikai cserét használtam a szakmai kollokáció miatt. |
| This format can be less memory efficient and make matrix algebra less easily applicable, but has the advantage of being able to add more variables (e.g., a sentiment score) and enables the use of the entire Tidyverse arsenal. | Ez a formátum kevésbé memóriahatékony lehet, és a mátrixalgebra kevésbé könnyen alkalmazható, de előnye, hogy több változót (pl. egy sentiment score-t) adhat hozzá, és lehetővé teszi a teljes Tidyverse-arzenál használatát. | Ez a formátum kevésbé memóriahatékony lehet, és a mátrixalgebra kevésbé könnyen alkalmazható, de előnye, hogy több változót (pl. egy érzelmi értékelést) adhat hozzá, és lehetővé teszi a teljes Tidyverse-arzenál használatát. | Ez a formátum kevésbé hatékony a memóriakezelés szempontjából, és a mátrixalgebra is nehezebben alkalmazható, de előnye, hogy több változót (pl. egy érzelmi értékelést) adhat hozzá a táblázathoz, és lehetővé teszi a teljes Tidyverse-eszköztár használatát. | A hatékony memóriakezelés érthetőbb. Az érzelmi értékelésről később lesz szó a szövegben. |
| Thus, for users that prefer the tidy data | Így a tidy adatfilozófiát kedvelő felhasználók | Így a tidy adatfilozófiát kedvelő felhasználók | Így a tidy adatkezelési elveket kedvelő | A filozófia kifejezést nagyképűnek, kissé |

| | | | | |
|---|---|---|---|---|
| philosophy, tidytext can be a good alternative package to quanteda or tm, although these packages can also be used together quite nicely depending on the particular operations desired. | számára a tidytext jó alternatív csomag lehet a quanteda vagy a tm mellett, bár ezek a csomagok a kívánt konkrét műveletektől függően együtt is nagyon szépen használhatók. | számára a tidytext jó alternatív program lehet a quanteda vagy a tm mellett, bár ezek a csomagok a kívánt konkrét műveletektől függően együtt is nagyon szépen használhatók. | felhasználók számára a tidytext jó alternatíva lehet a quanteda vagy a tm mellett, bár ezek a csomagok a kívánt konkrét műveletektől függően együtt is nagyon hatékonyan használhatók. | fellengzősnek találom, ezért módosítottam. |
| Consistent with the other examples in this teacher's corner, we demonstrate the creation of DTMs using the quanteda package. | A tanári sarok többi példájával összhangban a DTM-ek létrehozását a quanteda csomag segítségével mutatjuk be. | A módszertani segédlet többi példájával összhangban a DTM-ek létrehozását a quanteda csomag segítségével mutatjuk be. | A módszertani segédlet többi példájával összhangban a DTM-ek létrehozását a quanteda csomag segítségével mutatjuk be. | A gépi fordítás közel hibátlan. |
| Its dfm function provides a single line solution for creating a DTM from raw text, that also integrates the preprocessing techniques discussed above. | Ennek dfm függvénye egysoros megoldást nyújt a DTM létrehozására nyers szövegből, amely a fent tárgyalt előfeldolgozási technikákat is integrálja. | Ennek dfm funkciója egysoros megoldást nyújt a DTM létrehozására nyers szövegből, amely a fent tárgyalt előfeldolgozási technikákat is integrálja. | Ennek dfm funkciója a korábban bemutatott előfeldolgozási technikákat integrálva egyetlen lépésben nyújt megoldást a DTM létrehozására nyers szövegből. | A szórenden változtattam a gördülékenység növelése érdekében, de csak a teljes utószerkesztésnél. |
| These may also be built up through a sequence of lower-level functions, but many users find it convenient to go straight from a text or corpus to a DTM using this single function. | Ezeket alacsonyabb szintű függvények sorozatán keresztül is fel lehet építeni, de sok felhasználó számára kényelmes, ha egy szövegből vagy korpuszból közvetlenül ezzel az egyetlen függvénnyel jut el a DTM-hez. | Ezeket alacsonyabb szintű funkciók sorozatán keresztül is fel lehet építeni, de sok felhasználó számára kényelmes, ha egy szövegből vagy korpuszból közvetlenül ezzel az egyetlen lépéssel jut el a DTM-hez. | Ezeket alacsonyabb szintű funkciók sorozatán keresztül is fel lehet építeni, de sok felhasználó számára kényelmes, ha egy szövegből vagy korpuszból közvetlenül jut el a DTM-hez. | Kihagyást alkalmaztam |
| The DTM can also be created from a quanteda corpus object, which stores text and associated meta-data, including | A DTM létrehozható egy quanteda korpuszobjektumból is, amely a szöveget és a kapcsolódó metaadatokat | A DTM létrehozható egy quanteda korpuszobjektumból is, amely a szöveget és a kapcsolódó metaadatokat | A DTM egy quanteda korpuszobjektumból is létrehozható, amely a szöveget és a kapcsolódó metaadatokat tárolja, | |

| | | | | |
|---|---|---|---|---|
| document-level variables. | tárolja, beleértve a dokumentumszintű változókat is. | tárolja, beleértve a dokumentumszintű változókat is. | ideértve a dokumentumszintű változókat is. | |
| When a corpus is tokenized or converted into a DTM, these document-level variables are saved in the object, which can be very useful later when the documents in the DTM need to be used as covariates in supervised machine learning. | Amikor egy korpuszt tokenizálunk vagy DTM-mé alakítunk át, ezek a dokumentumszintű változók az objektumban kerülnek elmentésre, ami később nagyon hasznos lehet, amikor a DTM-ben lévő dokumentumokat kovariátorként kell használni a felügyelt gépi tanulásban. | Amikor egy korpuszt tokenizálunk vagy DTM-mé alakítunk át, ezek a dokumentumszintű változók az objektumban kerülnek elmentésre, ami később nagyon hasznos lehet, amikor a DTM-ben lévő dokumentumokat kovariátorként kell használni a felügyelt gépi tanulásban. | Amikor egy korpuszt tokenizálunk vagy DTM-mé alakítunk át, ezek a dokumentumszintű változókat az adott fájlba mentjük el.<br>Ez később nagyon hasznos lehet, amikor a DTM-ben lévő dokumentumokat kovariátorként (moderator vagy mediator változók) kell használnunk a felügyelt gépi tanulás során. | Az eredeti mondatot kettébontottam a jobb érthetőség miatt. A kovariátor-szó a magyarban kevésbé ismert, ezért kiegészítést tettem. |
| The stored document variables also make it possible to aggregate quanteda objects by groups, which is extremely useful when texts are stored in small units—like Tweets—but need to be aggregated in a DTM by grouping variables such as users, dates, or combinations of these. | A tárolt dokumentumváltozók lehetővé teszik a quanteda-objektumok csoportok szerinti aggregálását is, ami rendkívül hasznos, amikor a szövegek kis egységekben - például tweetekben - vannak tárolva, de a DTM-ben olyan változók csoportosításával kell aggregálni őket, mint a felhasználók, dátumok vagy ezek kombinációi. | A tárolt dokumentumváltozók lehetővé teszik a quanteda-objektumok csoportok szerinti aggregálását is, ami rendkívül hasznos, amikor a szövegek kis egységekben - például tweetekben - vannak tárolva, de a DTM-ben olyan változók csoportosításával kell aggregálni őket, mint például a felhasználók, a dátumok vagy ezek kombinációi. | A tárolt dokumentumváltozók lehetővé teszik a quanteda fájlok csoportok szerinti aggregálását is, ami rendkívül hasznos akkor, ha a szövegek kis egységekben - például tweetekben - vannak tárolva, de a DTM-ben olyan változók szerinti csoportosításban kell aggregálnunk őket, mint például a felhasználók, a dátumok vagy ezek kombinációi. | A gépi fordítás közel hibátlan, lexikai cserét alkalmaztam |
| Because quanteda is compatible with the | Mivel a quanteda kompatibilis a pkgreadtext | Mivel a quanteda kompatibilis a pkgreadtext | A quanteda kompatibilis a pkgreadtext programmal, a | Szórendi módosítás Mivel kihagyása. |

| | | | | |
|---|---|---|---|---|
| pkgreadtext package, creating a corpus from texts on disk takes only a single additional step. | csomaggal, a korpusz létrehozása a lemezen lévő szövegekből csak egyetlen további lépést igényel. | csomaggal, a korpusz létrehozása a tárolt szövegekből csak egyetlen további lépést igényel. | korpusz létrehozása a tárolt szövegekből csak egyetlen további lépést igényel. | |
| In the following example we create a DTM from the readtext data as imported above. | A következő példában a fentiek szerint importált readtext adatokból hozunk létre egy DTM-et. | A következő példában a fentiek szerint importált readtext adatokból hozunk létre egy DTM-et. | A következő példában a fentiek szerint importált readtext adatokból hozunk létre egy DTM-et. | A gépi fordítás kifogástalan. |
| Not all terms are equally informative for text analysis. | Nem minden kifejezés egyformán informatív a szövegelemzés szempontjából. | Nem minden kifejezés egyformán informatív a szövegelemzés szempontjából. | Nem minden kifejezés egyformán informatív a szövegelemzés szempontjából. | A gépi fordítás kifogástalan. |
| One way to deal with this is to remove these terms from the DTM. | Ezt úgy lehet kezelni, hogy ezeket a kifejezéseket eltávolítjuk a DTM-ből. | Ezt úgy lehet kezelni, hogy ezeket a kifejezéseket eltávolítjuk a DTM-ből. | Ezt úgy lehet kezelni, hogy ezeket a kifejezéseket eltávolítjuk a DTM-ből. | A gépi fordítás kifogástalan. |
| We have already discussed the use of stopword lists to remove very common terms, but there are likely still other common words and this will be different between corpora. | Már tárgyaltuk a stopword-listák használatát a nagyon gyakori kifejezések eltávolítására, de valószínűleg vannak még más gyakori szavak is, és ez az egyes testületek között eltérő lesz. | Már tárgyaltuk a stopword-listák használatát a nagyon gyakori kifejezések eltávolítására, de valószínűleg vannak még más gyakori szavak is, és ez az egyes testületek között eltérő lesz. | A nagyon gyakori kifejezések eltávolítására már tárgyaltuk a stopword-listák használatát, de általában vannak még más gyakori szavak is, arányuk viszont az az egyes korpuszok között eltérő lesz. | Szórendi cserét és lexikai cserét alkalmaztam a jobb érthetőség miatt. |
| Furthermore, it can be useful to remove very rare terms for tasks such as category prediction (Yang & Pedersen, 1997) or topic modeling (Griffiths & Steyvers, 2004). | Továbbá hasznos lehet a nagyon ritka kifejezések eltávolítása olyan feladatokhoz, mint a kategória-előrejelzés (Yang & Pedersen, 1997) vagy a témamodellezés (Griffiths & Steyvers, 2004). | Továbbá hasznos lehet a nagyon ritka kifejezések eltávolítása olyan feladatokhoz, mint a kategória-előrejelzés (Yang & Pedersen, 1997) vagy a témamodellezés (Griffiths & Steyvers, 2004). | A nagyon ritka kifejezések eltávolítása is hasznos lehet az olyan feladatokhoz, mint például az automatikus osztályzos (Yang & Pedersen, 1997) vagy a témamodellezés (Griffiths & Steyvers, 2004). | Olyan fordítást kerestem, ami a magyarban jobban érthetővé teszi az eredetit. |
| This is especially useful for improving efficiency, because it can greatly reduce the size of the | Ez különösen hasznos a hatékonyság javítása szempontjából, mivel nagymértékben | Ez különösen hasznos a hatékonyság javítása szempontjából, mivel nagymértékben | Ez különösen hasznos a hatékonyság javítása szempontjából, mert nagymértékben | Szórendi csere és az eredeti kifejezésetk tükörfordítása helyett kis magyarázat. |

| | | | | |
|---|---|---|---|---|
| vocabulary (i. e. , the number of unique terms), but it can also improve accuracy. | csökkentheti a szókincs méretét (azaz az egyedi kifejezések számát), de a pontosságot is javíthatja. | csökkentheti a szókincs méretét (azaz az egyedi kifejezések számát), de a pontosságot is javíthatja. | csökkentheti a korpuszban lévő egyedi szavak és kifejezések számát, sőt a pontosságot is javíthatja. | |
| A simple but effective method is to filter on document frequencies (the number of documents in which a term occurs), using a threshold for minimum and maximum number (or proportion) of documents (Griffiths & Steyvers, 2004; Yang & Pedersen, 1997). | Egy egyszerű, de hatékony módszer a dokumentumfrekvenciák (azon dokumentumok száma, amelyekben egy kifejezés előfordul) alapján történő szűrés, a dokumentumok minimális és maximális számának (vagy arányának) küszöbértékét használva (Griffiths & Steyvers, 2004; Yang & Pedersen, 1997). | Egy egyszerű, de hatékony módszer a dokumentumfrekvenciák (azon dokumentumok száma, amelyekben egy kifejezés előfordul) alapján történő szűrés, a dokumentumok minimális és maximális számának (vagy arányának) küszöbértékét használva (Griffiths & Steyvers, 2004; Yang & Pedersen, 1997). | Egyszerű, de hatékony módszer a dokumentumfrekvenciák (azon dokumentumok száma, amelyekben egy kifejezés előfordul) alapján történő szűrés, a dokumentumok minimális és maximális számának (vagy arányának) küszöbértékét használva (Griffiths & Steyvers, 2004; Yang & Pedersen, 1997). | A gépi fordítás kifogástalan. |
| Instead of removing less informative terms, an alternative approach is assign them variable weights. | A kevésbé informatív kifejezések eltávolítása helyett egy alternatív megközelítés az, hogy változó súlyokat rendelünk hozzájuk. | A kevésbé informatív kifejezések eltávolítása helyett egy alternatív megközelítés az, hogy változó súlyokat rendelünk hozzájuk. | A kevésbé informatív kifejezések eltávolítása helyett egy alternatív megközelítés lehet az, ha változó súlyokat rendelünk hozzájuk. | A gépi fordítás kifogástalan. |
| Many text analysis techniques perform better if terms are weighted to take an estimated information value into account, rather than directly using their occurrence frequency. | Számos szövegelemzési technika jobban teljesít, ha a kifejezések súlyozása egy becsült információérték figyelembevételével történik, ahelyett, hogy közvetlenül az előfordulási gyakoriságukat használnánk. | Számos szövegelemzési technika jobban teljesít, ha a kifejezések súlyozása egy becsült információérték figyelembevételével történik, ahelyett, hogy közvetlenül az előfordulási gyakoriságukat használnánk. | Számos szövegelemzési eljárás jobban teljesít, ha a kifejezések súlyozása egy becsült információérték figyelembevételével történik, ahelyett, hogy közvetlenül az előfordulási gyakoriságukat használnánk. | A gépi fordítás kifogástalan. |
| Given a sufficiently large corpus, we can use information about the | Kellően nagy korpusz esetén a kifejezések eloszlásáról szóló | Kellően nagy korpusz esetén a kifejezések eloszlásáról szóló | Kellően nagy korpusz esetén a kifejezések eloszlásával kapcsolatos | Lexikai cserét alkalmaztam |

| | | | | |
|---|---|---|---|---|
| distribution of terms in the corpus to estimate this information value. | információkat felhasználhatjuk a korpuszban az információs érték becsléséhez. | információkat felhasználhatjuk a korpuszban az információs érték becsléséhez. | információkat felhasználhatjuk a korpuszban az információs érték becsléséhez. | |
| A popular weighting scheme that does so is term frequency-inverse document frequency (tf-idf), which down-weights that occur in many documents in the corpus. | Egy népszerű súlyozási séma, amely ezt teszi, a terminusfrekvencia-inverz dokumentumfrekvencia (tf-idf), amely a korpuszban sok dokumentumban előforduló kifejezéseket lefelé súlyozza. | Egy népszerű súlyozási séma, amely ezt teszi, a terminusfrekvencia-inverz dokumentumfrekvencia (tf-idf), amely a korpuszban sok dokumentumban előforduló kifejezéseket lefelé súlyozza. | A terminusfrekvencia-inverz dokumentumfrekvencia (tf-idf) egy olyan elterjedt súlyozási séma, amely a korpusz sok dokumentumában előforduló kifejezéseket kisebb súlyszám-értékkel veszi figyelembe. | Szórendi csere |
| Using a document frequency threshold and weighting can easily be performed on a DTM. | A dokumentumfrekvencia küszöbérték és súlyozás használata könnyen elvégezhető a DTM-en. | A dokumentumfrekvencia küszöbérték és súlyozás használata könnyen elvégezhető a DTM-en. | A dokumentumfrekvencia küszöbérték és súlyozás használata egyszerűen elvégezhető a DTM-en. | A gépi fordítás gyakorlatilag kifogástalan, apró lexikai csere, stiláris okokból. |
| quanteda includes the functions docfreq, tf, and tfidf, for obtaining document frequency, term frequency, and tf-idf respectively. | A quanteda tartalmazza a docfreq, tf és tfidf függvényeket a dokumentumfrekvencia, a terminusfrekvencia és a tf-idf kiszámításához. | A quanteda tartalmazza a docfreq, tf és tfidf funkciókat a dokumentumfrekvencia, a terminusfrekvencia és a tf-idf kiszámításához. | A quanteda tartalmazza a docfreq, tf és tfidf funkciókat a dokumentumfrekvencia, a terminusfrekvencia és a tf-idf kiszámításához. | A gépi fordítás gyakorlatilag kifogástalan. |
| Each function has numerous options for implementing the SMART weighting scheme (Manning et al. , 2008). | Mindegyik függvény számos lehetőséggel rendelkezik a SMART súlyozási séma megvalósításához (Manning et al. , 2008). | Mindegyik funkció számos lehetőséggel rendelkezik a SMART súlyozási séma megvalósításához (Manning et al.,2008). | Mindegyik funkció számos lehetőségget kínál a SMART súlyozási séma megvalósításához (Manning et al., 2008) | A gépi fordítás gyakorlatilag kifogástalan, de a függvényt funkcióra módosítottam. |
| As a high-level wrapper to these, quanteda also provides the dfm_weight function. | A quanteda ezekhez magas szintű burkolóként a dfm_weight függvényt is biztosítja. | A quanteda ezekhez átfogó, magas szintű lehetőségként a dfm_weight funkciót is biztosítja. | A quanteda ezekhez átfogó, magas szintű lehetőségként a dfm_weight funkciót is biztosítja. | Lexikai cserét alkalmaztam |

| | | | | |
|---|---|---|---|---|
| In the example below, the word “senat[e]” has a higher weight than the less informative term “among”, which both occur once in the first document. | Az alábbi példában a "senat[e]" szónak nagyobb súlya van, mint a kevésbé informatív "among" kifejezésnek, amelyek mindketten egyszer fordulnak elő az első dokumentumban. | Az alábbi példában a "senat[e]" szónak nagyobb súlya van, mint a kevésbé informatív "among" kifejezésnek, amelyek egyszer-egyszer fordulnak elő az első dokumentumban. | Az alábbi példában a "senat[e]" szónak nagyobb súlya van, mint a kevésbé informatív "among" kifejezésnek, amelyek egyszer-egyszer fordulnak elő az első dokumentumban. | A gépi fordítás gyakorlatilag kifogástalan. |
| For an overview of text analysis approaches we build on the classification proposed by Boumans and Trilling (2016) in which three approaches are distinguished: counting and dictionary methods, supervised machine learning, and unsupervised machine learning. | A szövegelemzési megközelítések áttekintéséhez a Boumans és Trilling (2016) által javasolt osztályozásra építünk, amelyben három megközelítést különböztetünk meg: számláló és szótári módszerek, felügyelt gépi tanulás és felügyelet nélküli gépi tanulás. | A szövegelemzési megközelítések áttekintéséhez a Boumans és Trilling (2016) által javasolt osztályozásra építünk, amelyben három módszert különböztetünk meg: számláló és szótári módszerek, felügyelt gépi tanulás és felügyelet nélküli gépi tanulás. | A szövegelemzési megközelítések áttekintéséhez a Boumans és Trilling (2016) által javasolt osztályozásra építünk, amelyben három módszert különböztet meg: számláló és szótári módszerek, felügyelt gépi tanulás és felügyelet nélküli gépi tanulás. | Lexikai csere |
| They position these approaches, in this order, on a dimension from most deductive to most inductive. | Ezeket a megközelítéseket ebben a sorrendben a leginkább deduktívtól a leginkább induktívig terjedő dimenzióban helyezik el. | Ezeket a megközelítéseket ebben a sorrendben a leginkább deduktívtól a leginkább induktívig terjedő dimenzióban helyehetjük el. | Ezek a módszereket ebben a sorrendben a leginkább deduktívtól a leginkább induktívig terjedő koordinátatengely mentén helyezkednek el. | Lexikai csere |
| Deductive, in this scenario, refers to the use of an a priori defined coding scheme. | A deduktív ebben a forgatókönyvben egy a priori meghatározott kódolási séma használatára utal. | A deduktív ebben az esetben egy előre (a priori) meghatározott kódolási séma használatára utal. | A deduktív ebben az esetben egy előre (a priori) meghatározott kódolási séma használatára utal. | Kiegészítést és lexikai cserét alkalmaztam |
| In other words, the researchers know beforehand what they are looking for, and only | Más szóval a kutatók előre tudják, hogy mit keresnek, és csak az elemzés | Más szóval a kutatók előre tudják, hogy mit keresnek, és csak az elemzés | Más szóval a kutatók előre tudják, hogy mit keresnek, és csak az elemzés | A gépi fordítás gyakorlatilag kifogástalan. |

| | | | | |
|---|---|---|---|---|
| seek to automate this analysis. | automatizálására törekszenek. | automatizálására törekszenek. | automatizálására törekszenek. | |
| The relation to the concept of deductive reasoning is that the researcher assumes that certain rules, or premises, are true (e.g. , a list of words that indicates positive sentiment) and thus can be applied to draw conclusions about texts. | A kapcsolat a deduktív érvelés fogalmával az, hogy a kutató feltételezi, hogy bizonyos szabályok vagy premisszák igazak (pl. a pozitív hangulatot jelző szavak listája), és így alkalmazhatók a szövegekre vonatkozó következtetések levonására. | A kapcsolat a deduktív érvelés fogalmával az, hogy a kutató feltételezi, hogy bizonyos szabályok vagy premisszák igazak (pl. a pozitív hangulatot jelző szavak listája), és így alkalmazhatók a szövegekre vonatkozó következtetések levonására. | A kapcsolat a deduktív érvelés fogalmával az, hogy a kutató feltételezi, hogy bizonyos szabályok vagy premisszák igazak (pl. a pozitív hangulatot jelző szavak listája), és így alkalmazhatók a szövegekre vonatkozó következtetések levonására. | A gépi fordítás gyakorlatilag kifogástalan. |
| Inductive, in contrast, means here that instead of using an a priori coding scheme, the computer algorithm itself somehow extracts meaningful codes from texts. | Az induktív ezzel szemben itt azt jelenti, hogy egy a priori kódolási séma használata helyett a számítógépes algoritmus maga vonja ki valahogyan az értelmes kódokat a szövegekből. | Az induktív ezzel szemben itt azt jelenti, hogy egy a priori kódolási séma használata helyett a számítógépes algoritmus maga keresi meg az értelmes kódokat a szövegekből. | Az induktív ezzel szemben itt azt jelenti, hogy egy a priori kódolási séma használata helyett a számítógépes algoritmus maga határozza meg az értelmes kódokat a szövegekből. | A gépi fordítás gyakorlatilag kifogástalan. |
| For example, by looking for patterns in the co-occurrence of words and finding latent factors (e. g. , topics, frames, authors) that explain these patterns—at least mathematically. | Például úgy, hogy mintákat keres a szavak együttes előfordulásában, és olyan látens tényezőket (pl. témák, keretek, szerzők) talál, amelyek ezeket a mintákat - legalábbis matematikailag - megmagyarázzák. | Például úgy, hogy mintákat keres a szavak együttes előfordulásában, és olyan látens tényezőket témák, keretek, szerzők) talál, amelyek ezeket a mintákat - legalábbis matematikailag - megmagyarázzák. | Ez például úgy történhet, hogy mintázatokat keres a szavak együttes előfordulásában, és olyan látens tényezőket (pl. témák, keretek, szerzők) talál, amelyek ezeket a mintákat - legalábbis matematikailag - megmagyarázzák. | Kiegészítést alkalmaztam a stílus javítására, de csak a teljes utószerkesztésnél. |
| In terms of inductive reasoning, it can be said that the algorithm creates broad generalizations | Az induktív érvelés szempontjából azt mondhatjuk, hogy az algoritmus konkrét megfigyelések alapján | Az induktív érvelés szempontjából azt mondhatjuk, hogy az algoritmus konkrét megfigyelések alapján, | Az induktív érvelés szempontjából azt mondhatjuk, hogy az algoritmus konkrét megfigyelések alapján | A gépi fordítás gyakorlatilag kifogástalan. |

| | | | | |
|---|---|---|---|---|
| based on specific observations. | széleskörű általánosításokat hoz létre. | széleskörű általánosításokat hoz létre. | széleskörű általánosításokat hoz létre. | |
| In addition to these three categories, we also consider a statistics category, encompassing all techniques for describing a text or corpus in numbers. | E három kategória mellett egy statisztikai kategóriát is figyelembe veszünk, amely magában foglalja a szöveg vagy korpusz számokkal való leírására szolgáló összes technikát. | E három kategória mellett egy statisztikai kategóriát is figyelembe veszünk, amely magában foglalja a szöveg vagy korpusz számokkal való leírására szolgáló összes technikát. | E három kategória mellett egy statisztikai kategóriát is figyelembe veszünk, amely magában foglalja a szöveg vagy korpusz számokkal történő leírására szolgáló összes technikát. | A gépi fordítás gyakorlatilag kifogástalan. |
| Like unsupervised learning, these techniques are inductive in the sense that no a priori coding scheme is used, but they do not use machine learning. | A felügyelet nélküli tanuláshoz hasonlóan ezek a technikák is induktívak abban az értelemben, hogy nem használnak a priori kódolási sémát, de nem alkalmaznak gépi tanulást. | A felügyelet nélküli tanuláshoz hasonlóan ezek a technikák is induktívak abban az értelemben, hogy nem használnak a priori kódolási sémát, de nem alkalmaznak gépi tanulást. | A felügyelet nélküli tanuláshoz hasonlóan ezek a technikák is induktívak abban az értelemben, hogy sem a priori kódolási sémát, sem gépi tanulást nem alkalmaznak. | Leegyszerűsítettem, átfogalmaztam a mondatot, de csak a teljes utószerkesztésnél.. |
| For the example code for each type of analysis, we use the Inaugural Addresses of US presidents (N = 58) that is included in the quanteda package. | Az egyes elemzési típusok példakódjához az amerikai elnökök beiktatási beszédeit (N = 58) használjuk, amely a quanteda csomagban szerepel. | Az egyes elemzési típusok példakódjához az amerikai elnökök beiktatási beszédeit (N = 58) használjuk, amely a quanteda csomagban szerepel. | Az egyes elemzési típusokat bemutató minta-parancsokhoz az amerikai elnökök- a quanteda programban szereplő-beiktatási beszédeit (N = 58) használjuk. | Lexikai cserét hajtottam végre a jobb érthetőség érdekében |
| | | | | |
| The dictionary approach broadly refers to the use of patterns—from simple keywords to complex Boolean queries and regular expressions—to count how often certain concepts occur in texts. | A szótári megközelítés tágabb értelemben a minták használatára utal - az egyszerű kulcsszavaktól az összetett Boolean-lekérdezésekig és a szabályos kifejezésekig -, hogy megszámoljuk, milyen gyakran fordulnak elő bizonyos fogalmak a szövegekben. | A szótári megközelítés tágabb értelemben a minták használatára utal - az egyszerű kulcsszavaktól az összetett Bool-algebrai-lekérdezésekig és a szabályos kifejezésekig -, annak érdekében, hogy megszámoljuk, milyen gyakran fordulnak elő bizonyos fogalmak a szövegekben. | A tágabb értelemben vett szótári megközelítés a mintázatok felhasználására utal - az egyszerű kulcsszavaktól az összetett Bool-algebrai-lekérdezésekig és a kötött kifejezésekig, állandósult szókapcsolatokig –, annak érdekében, hogy megszámoljuk, milyen gyakran fordulnak elő | Kiegészítést és lexikai cserét alkalmaztam, valamint lecseréletm a kiskötőjelet gondolatjelre, de csak a teljes utószerkesztésnél. |

| | | | | |
|---|---|---|---|---|
| | | | bizonyos fogalmak a szövegekben. | |
| This is a deductive approach, because the dictionary defines a priori what codes are measured and how, and this is not affected by the data. | Ez egy deduktív megközelítés, mivel a szótár a priori meghatározza, hogy milyen kódokat és hogyan kell mérni, és ezt az adatok nem befolyásolják. | Ez deduktív megközelítés, mivel a szótár a priori meghatározza, hogy milyen kódokat és hogyan kell mérni, és ezt az adatok nem befolyásolják. | Ez deduktív megközelítés, mert a szótár a priori meghatározza, hogy milyen kódokat és hogyan kell mérni, és ezt az adatok nem befolyásolják. | Kihagyást alkalmaztam,kihagytam az egy határozatlan névelőt stilisztikai szempontból. |
| 8 Using dictionaries is a computationally simple but powerful approach. | 8 A szótárak használata számítási szempontból egyszerű, de hatékony megközelítés. | A szótárak használata számítási szempontból egyszerű, de hatékony megközelítés. | A szótárak használata számítási szempontból egyszerű, de hatékony módszer. | Lexikai cserét alkalmaztam |
| It has been used to study subjects such as media attention for political actors (Schuck, Xezonakis, Elenbaas, Banducci, & De Vreese, 2011; Vliegenthart, Boomgaarden, & Van Spanje, 2012) and framing in corporate news (Schultz, Kleinnijenhuis, Oegema, Utz, & Van Atteveldt, 2012). | Olyan témák tanulmányozására használták, mint a politikai szereplőkre irányuló médiafigyelem (Schuck, Xezonakis, Elenbaas, Banducci és De Vreese, 2011; Vliegenthart, Boomgaarden és Van Spanje, 2012) és a vállalati hírek keretezése (Schultz, Kleinnijenhuis, Oegema, Utz és Van Atteveldt, 2012). | Olyan témák tanulmányozására használták, mint például a politikai szereplőkre irányuló médiafigyelem (Schuck, Xezonakis, Elenbaas, Banducci és De Vreese, 2011; Vliegenthart, Boomgaarden és Van Spanje, 2012), vagy a vállalati információk kontextusba foglalása (Schultz, Kleinnijenhuis, Oegema, Utz és Van Atteveldt, 2012). | Olyan témák tanulmányozására használták, mint például a politikai szereplőkre irányuló médiafigyelem (Schuck, Xezonakis, Elenbaas, Banducci és De Vreese, 2011; Vliegenthart, Boomgaarden és Van Spanje, 2012), vagy a vállalati információk kontextusba foglalása (Schultz, Kleinnijenhuis, Oegema, Utz és Van Atteveldt, 2012). | Mind a részleges, mind a teljes utószerkesztésénél tartalmi javítást végeztem, mert az eredeti angol szövegből nem derül ki az, hogy itt mi volt a kutatás célja. A hivatkozott cikk viszont azt mutatja be, hogy egy környzetszennyezési botrány után milyen kontextusba helyezte a kommunikációját egy olajvállalat. |
| Dictionaries are also a popular approach for measuring sentiment (De Smedt & Daelemans, 2012; Mostafa, 2013; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) as well as other | A szótárak szintén népszerű megközelítésnek számítanak a szentiment mérésére (De Smedt & Daelemans, 2012; Mostafa, 2013; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011), valamint a | A szótárak szintén népszerű megközelítésnek számítanak az érzelmek mérésére (De Smedt & Daelemans, 2012; Mostafa, 2013; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011), valamint a | A szótárak szintén népszerű eszköznek számítanak az érzelmek (De Smedt & Daelemans, 2012; Mostafa, 2013; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011), valamint a | A gépi fordítás gyakorlatilag kifogástalan. |

| | | | | |
|---|---|---|---|---|
| dimensions of subjective language (Tausczik & Pennebaker, 2010). | szubjektív nyelv más dimenzióinak mérésére (Tausczik & Pennebaker, 2010). | szubjektív nyelvhasználat más dimenzióinak mérésére (Tausczik & Pennebaker, 2010). | szubjektív nyelvhasználat más dimenzióinak mérésére (Tausczik & Pennebaker, 2010). | |
| By combining this type of analysis with information from advanced NLP techniques for identifying syntactic clauses, it also becomes possible to perform more fine-grained analyses, such as sentiment expressions attributed to specific actors (Van Atteveldt, 2008), or actions and affections from one actor directed to another (Van Atteveldt, Sheafer, Shenhav, & Fogel-Dror, 2017). | Az ilyen típusú elemzések és a szintaktikai záradékok azonosítására szolgáló fejlett NLP-technikákból származó információk kombinálásával finomabb szemcseméretű elemzések elvégzése is lehetővé válik, például az egyes szereplőknek tulajdonított érzelemkifejezések (Van Atteveldt, 2008) vagy az egyik szereplőtől a másiknak címzett cselekvések és érzelmek (Van Atteveldt, Sheafer, Shenhav, & Fogel-Dror, 2017). | Az ilyen típusú elemzések és a szintaktikai egységek azonosítására szolgáló fejlett NLP-technikákból származó információk kombinálásával részletesebb elemzések elvégzése is lehetővé válik, például az egyes szereplőknek tulajdonított érzelemkifejezések (Van Atteveldt, 2008) vagy az egyik szereplőtől a másik felé irányuló cselekvések és érzelmek (Van Atteveldt, Sheafer, Shenhav, & Fogel-Dror, 2017). | Az ilyen típusú elemzések és a szintaktikai egységek azonosítására szolgáló fejlett NLP-technikákból származó információk kombinálásával részletesebb elemzések elvégzése is lehetővé válik, például az egyes szereplőknek tulajdonított érzelemkifejezések (Van Atteveldt, 2008) vagy az egyik szereplőtől a másik felé irányuló cselekvések és érzelmek vizsgálata (Van Atteveldt, Sheafer, Shenhav, & Fogel-Dror, 2017). | Lexikai csere, mert a mechanikus, tükörfordítás fordítás félreérthető lenne. |
| The following example shows how to apply a dictionary to a quanteda DTM. | A következő példa azt mutatja be, hogyan alkalmazhat egy szótárat egy quanteda DTM-re. | A következő példa azt mutatja be, hogyan alkalmazhat egy szótárat egy quanteda DTM-re. | A következő példa azt mutatja be, hogyan alkalmazható egy szótár egy quanteda DTM-re. | A gépi fordítás gyakorlatilag kifogástalan. |
| The first step is to create a dictionary object (here called myDict), using the dictionary function. | Az első lépés egy szótár objektum létrehozása (itt myDict), a dictionary függvény használatával. | Az első lépés egy szótár fájl létrehozása (itt myDict), a dictionary függvény használatával. | Az első lépés egy szótár fájl létrehozása (itt myDict), a dictionary funkció használatával. | Lexikai csere az egyértelműsítés érdekében. |
| For simplicity our example uses a very simple dictionary, but it is is also possible to import large, pre-made | Az egyszerűség kedvéért példánk egy nagyon egyszerű szótárat használ, de lehetőség van nagy, előre elkészített szótárak | Az egyszerűség kedvéért példánk egy nagyon egyszerű szótárat használ, de lehetőség van nagy, előre elkészített szótárak | Az egyszerűség kedvéért példánk egy nagyon egyszerű szótárat használ, de lehetőség van nagy, előre elkészített szótárak | A gépi fordítás gyakorlatilag kifogástalan. |

| | | | | |
|---|---|---|---|---|
| dictionaries, including files in other text analysis dictionary formats such as LIWC, Wordstat, and Lexicoder. | importálására is, beleértve más szövegelemző szótárformátumú fájlokat is, mint például a LIWC, Wordstat és Lexicoder. | importálására is, beleértve más szövegelemző szótárformátumú fájlokat is, mint például a LIWC, Wordstat és Lexicoder. | importálására is, ideértve más szövegelemző szótár-formátumú fájlokat is, mint például a LIWC, Wordstat és a Lexicoder. | |
| Dictionaries can also be written and imported from YAML files, and can include patterns of fixed matches, regular expressions, or the simpler “glob” pattern match (using just * and ? for wildcard characters) common in many dictionary formats. | A szótárak YAML fájlokból is írhatók és importálhatók, és tartalmazhatnak fix egyezésekből, reguláris kifejezésekből vagy az egyszerűbb "glob" mintaillesztésből álló mintákat (csak * és ? karaktereket használva joker karakterekként), amelyek számos szótárformátumban elterjedtek. | A szótárak YAML fájlokból is írhatók és importálhatók, és tartalmazhatnak fix egyezésekből, szabályos kifejezésekből vagy az egyszerűbb "glob" mintaillesztésből álló mintákat (csak * és ? karaktereket használva joker karakterekként), amelyek számos szótárformátumban elterjedtek. | A szótárak YAML fájlokból is írhatók és importálhatók, és tartalmazhatnak fix egyezésekből, szabályos kifejezésekből vagy az egyszerűbb "glob" szerkezetekből álló mintázatokat, (csak * és ? karaktereket használva joker karakterekként), amelyek számos szótárformátumban elterjedtek. | A gépi fordítás gyakorlatilag kifogástalan. |
| With the dfm_lookup function, the dictionary object can then be applied on a DTM to create a new DTM in which columns represent the dictionary codes. | A dfm_lookup függvénnyel a szótárobjektumot ezután egy DTM-re lehet alkalmazni egy új DTM létrehozásához, amelyben az oszlopok a szótárkódokat képviselik. | A dfm_lookup függvénnyel a szótárobjektumot ezután egy DTM-re lehet alkalmazni egy új DTM létrehozásához, amelyben az oszlopok a szótárkódokat képviselik. | A dfm_lookup funkcióval a szótárobjektumot egy DTM-re lehet alkalmazni egy új DTM létrehozásához, amelyben a szótárkódokat az oszlopok tartalmazzák. | Szórendi módosítás és lexikai csere |
| The supervised machine learning approach refers to all classes of techniques in which an algorithm learns patterns from an annotated set of training data. | A felügyelt gépi tanulási megközelítés a technikák minden olyan osztályára vonatkozik, amelyben egy algoritmus mintákat tanul egy megjegyzésekkel ellátott gyakorlóadat-halmazból. | A felügyelt gépi tanulás módszere a technikák minden olyan osztályára vonatkozik, amelyben egy algoritmus mintákat tanul egy további információkkal kiegészített gyakorló adathalmazból. | A felügyelt gépi tanulás módszere minen olyan kutatási technikát magába foglal, amikor egy algoritmus mintákat tanul egy további információkkal kiegészített gyakorló adathalmazból. | Lexikai csere, betoldás |

| | | | | |
|---|---|---|---|---|
| The intuitive idea is that these algorithms can learn how to code texts if we give them enough examples of how they should be coded. | Az intuitív gondolat az, hogy ezek az algoritmusok megtanulják, hogyan kell kódolni a szövegeket, ha elegendő példát adunk nekik arra, hogyan kell kódolni őket. | Az intuitív gondolat az, hogy ezek az algoritmusok megtanulják, hogyan kell kódolni a szövegeket, ha elegendő példát adunk nekik arra, hogyan kell kódolni őket. | Az intuitív módszer úgy írható le, hogy ezek az algoritmusok megtanulják, hogyan kell kódolni a szövegeket, ha arra elegendő példát adunk nekik. | Lexikai csere |
| A straightforward example is sentiment analysis, using a set of texts that are manually coded as positive, neutral, or negative, based on which the algorithm can learn which features (words or word combinations) are more likely to occur in positive or negative texts. | Egyszerű példa erre a hangulatelemzés, amely egy kézzel pozitívnak, semlegesnek vagy negatívnak kódolt szövegkészletet használ, amely alapján az algoritmus megtanulhatja, hogy mely jellemzők (szavak vagy szóösszetételek) fordulnak elő nagyobb valószínűséggel pozitív vagy negatív szövegekben. | Egyszerű példa erre a hangulatelemzés, amely egy ember által pozitívnak, semlegesnek vagy negatívnak kódolt szövegkészletet használ, amely alapján az algoritmus megtanulhatja, hogy mely jellemzők (szavak vagy szóösszetételek) fordulnak elő nagyobb valószínűséggel pozitív vagy negatív szövegekben. | Egyszerű példa erre a hangulatelemzés, amely egy ember által pozitívnak, semlegesnek vagy negatívnak kódolt szövegkészletet használ. Ennek alapján az algoritmus megtanulhatja, hogy mely jellemzők (szavak vagy szóösszetételek) fordulnak elő nagyobb valószínűséggel pozitív vagy negatív szövegekben. | A gépi fordítás gyakorlatilag kifogástalan. |
| Given an unseen text (from which the algorithm was not trained), the sentiment of the text can then be estimated based on the presence of these features. | Egy nem látott szöveg (amelyből az algoritmust nem képezték ki) alapján a szöveg hangulatát ezután ezeknek a jellemzőknek a jelenléte alapján lehet megbecsülni. | Egy ismeretlen szöveg (amelyre az algoritmust nem tanították be) alapján a szöveg hangulatát ezután ezeknek a jellemzőknek a jelenléte alapján lehet megbecsülni. | Egy ismeretlen szöveg segítségével (amelyre az algoritmust nem tanították be) a szöveg hangulatát ezeknek a jellemzőknek a jelenléte alapján meg lehet becsülni. | Szórendi csere, a helyes mondatfókusz miatt. |
| The deductive part is that the researchers provide the training data, which contains good examples representing the categories that the researchers are | A deduktív rész az, hogy a kutatók biztosítják a képzési adatokat, amelyek jó példákat tartalmaznak, amelyek reprezentálják azokat a kategóriákat, amelyeket a kutatók | A deduktív rész az, hogy a kutatók biztosítják a tanulási adatokat, amelyek olyan jó példákat tartalmaznak, amelyek reprezentálják azokat a kategóriákat, amiket a | A deduktív fázisban a kutatók biztosítják a tanuláshoz szükséges jó példákat tartalmazó adatokat, kiegészítve azokat a megfelelő kategóriákkal. | Az eredeti szöveg kissé zavaros, az a szakmailag és a tartalom szempontjából értelmes, ami a teljesen átszerkesztett verzióban található. |

| | | | | |
|---|---|---|---|---|
| attempting to predict or measure. | megpróbálnak megjósolni vagy mérni. | kutatók megpróbálnak előre jelezni vagy mérni. | | |
| However, the researchers do not provide explicit rules for how to look for these codes. | A kutatók azonban nem adnak explicit szabályokat arra vonatkozóan, hogy hogyan keressük ezeket a kódokat. | A kutatók azonban nem adnak explicit szabályokat arra vonatkozóan, hogy hogyan keressük ezeket a kategóriákat. | A kutatók azonban nem adnak explicit szabályokat arra vonatkozóan, hogy hogyan keressük ezeket a kategóriákat. | Lexikai cserét hajtottam végre, egyébként értelmetlen a szöveg. |
| The inductive part is that the supervised machine learning algorithm learns these rules from the training data. | Az induktív rész az, hogy a felügyelt gépi tanulási algoritmus megtanulja ezeket a szabályokat a képzési adatokból. | Az induktív rész az, hogy a felügyelt gépi tanulási algoritmus megtanulja ezeket a szabályokat a tanítására felhasznált adatokból. | Az induktív szakaszban a felügyelt gépi tanulási algoritmus a tanítására felhasznált adatokból tanulja meg ezeket a szabályokat. | Szórendi és lexikai csere |
| To paraphrase a classic syllogism: if the training data is a list of people that are either mortal or immortal, then the algorithm will learn that all men are extremely likely to be mortal, and thus would estimate that Socrates is mortal as well. | Egy klasszikus szillogizmust parafrazálva: ha a képzési adatok olyan emberek listája, akik vagy halandóak, vagy halhatatlanok, akkor az algoritmus megtanulja, hogy minden ember nagy valószínűséggel halandó, és így azt is megbecsülné, hogy Szókratész is halandó. | Egy klasszikus szillogizmust parafrazálva: ha a képzési adatok olyan emberek listája, akik vagy halandóak, vagy halhatatlanok, akkor az algoritmus megtanulja, hogy minden ember nagy valószínűséggel halandó, és így azt is megbecsülné, hogy Szókratész is halandó. | Egy klasszikus szillogizmust parafrazálva: ha a képzési adatok olyan emberek listája, akik vagy halandóak, vagy halhatatlanok, akkor az algoritmus megtanulja, hogy minden ember nagy valószínűséggel halandó, és így azt is megállapítaná, hogy Szókratész is halandó. | A gépi fordítás gyakorlatilag kifogástalan. |
| To demonstrate this, we train a model to predict whether an Inaugural Address was given before World War II—which we expect because prominent issues shift over time, and after wars in particular. | Ennek demonstrálására egy modellt képezünk ki annak előrejelzésére, hogy egy beiktatási beszédet a II. világháború előtt mondtak-e - amire azért számítunk, mert a kiemelkedő témák idővel, különösen a háborúk után változnak. | Ennek demonstrálására egy modellt képezünk ki annak előrejelzésére, hogy egy beiktatási beszédet a II. Világháború előtt mondták-e - el amire azért számítunk, mert a kiemelt témák idővel, különösen a háborúk után változnak. | Ennek bemutatására egy modellt tanítottunk be annak előrejelzésére, hogy egy [amerikai elnöki] beiktatási beszédet a II. világháború előtt vagy után mondtak-e el, aminek sikerességére azért számítunk, mert – különösen a világháború után – változtak a kiemelt témák idővel. | Beszúrást alkalmaztam a magyar olvasó tájékoztatására. Az eredeti angol szövegben az “after wars” kifejezés értelmetlen, itt egy háborúról van szó. második világháború vagy II. világháború (MHSz. 356.) |

| | | | | |
|---|---|---|---|---|
| | | | | Gondolatjere javítottam a kiskötőjelet. |
| Some dedicated packages for supervised machine learning are RTextTools (Jurka, Collingwood, Boydstun, Grossman, & Van Atteveldt, 2014) and kerasR (Arnold, 2017b). | A felügyelt gépi tanuláshoz néhány dedikált csomag az RTextTools (Jurka, Collingwood, Boydstun, Grossman, & Van Atteveldt, 2014) és a kerasR (Arnold, 2017b). | A felügyelt gépi tanuláshoz többféle program áll rendelkezésre, pl. az RTextTools (Jurka, Collingwood, Boydstun, Grossman, & Van Atteveldt, 2014) és a kerasR (Arnold, 2017b). | A felügyelt gépi tanuláshoz többféle program áll rendelkezésre, pl. az RTextTools (Jurka, Collingwood, Boydstun, Grossman, & Van Atteveldt, 2014) és a kerasR (Arnold, 2017b). | A gépi fordítás gyakorlatilag kifogástalan. |
| For this example, however, we use a classifier that is included in quanteda. | Ehhez a példához azonban egy olyan osztályozót használunk, amely a quanteda programban szerepel. | Ehhez a példához azonban egy olyan oszályozó algoritmust használunk, amely a quanteda programban szerepel. | Ehhez a példához a quanteda egyik oszályozó algoritmusát használunk. | Lexikai és szórendi csere |
| Before we start, we set a custom seed for R's random number generator so that the results of the random parts of the code are always the same. | Mielőtt elkezdenénk, beállítunk egy egyéni magot az R véletlenszám-generátorának, hogy a kód véletlenszerű részeinek eredményei mindig ugyanazok legyenek. | Mielőtt elkezdenénk, beállítunk egy egyéni kiinduló értéket az R véletlenszám-generátorának, hogy a kód véletlenszerű részeinek eredményei mindig ugyanazok legyenek. | Mielőtt elkezdenénk, beállítunk egy egyéni kiinduló értéket az R véletlenszám-generátorának, hogy a kód véletlenszerű részeinek eredményei mindig ugyanazok legyenek. | A gépi fordítás gyakorlatilag kifogástalan. |
| To prepare the data, we add the document (meta) variable is_prewar to the DTM that indicates which documents predate 1945. | Az adatok előkészítéséhez hozzáadjuk a DTM-hez az is_prewar dokumentum (meta) változót, amely jelzi, hogy mely dokumentumok vannak 1945 előtt. | Az adatok előkészítéséhez hozzáadjuk a DTM-hez az is_prewar dokumentum (meta) változót, amely jelzi, hogy mely dokumentumok vannak 1945 előtt. | Az adatok előkészítéséhez hozzáadjuk a DTM-hez az is_prewar dokumentum (meta) változót, amely jelzi, hogy mely dokumentumok származnak 1945 előttről. | A gépi fordítás gyakorlatilag kifogástalan. |
| This is the variable that our model will try to predict. | Ez az a változó, amelyet a modellünk megpróbál megjósolni. | Ez az a változó, amelyet a modellünk megpróbál előre jelezni. | Ez az a változó, amelyet a modellünk megpróbál előre jelezni. | Lexikai csere: a "megjósolni" ige ebben |

| | | | | |
|---|---|---|---|---|
| | | | | az összefüggéeben vulgáris. |
| We then split the DTM into training (train_dtm) and test (test_dtm) data, using a random sample of 40 documents for training and the remaining 18 documents for testing. | Ezután a DTM-et képzési (train_dtm) és teszt (test_dtm) adatokra osztjuk, a képzéshez 40 dokumentumból álló véletlenszerű mintát, a teszteléshez pedig a maradék 18 dokumentumot használjuk. | Ezután a DTM-et tanulási (train_dtm) és teszt (test_dtm) adatokra osztjuk, a tanuláshoz 40 dokumentumból álló véletlenszerű mintát, a teszteléshez pedig a maradék 18 dokumentumot használjuk. | Ezután a DTM-et tanulási (train_dtm) és teszt (test_dtm) adatokra osztjuk, a tanuláshoz 40 dokumentumból álló véletlenszerű mintát, a teszteléshez pedig a maradék 18 dokumentumot használjuk. | A gépi fordítás gyakorlatilag kifogástalan. |
| The training data is used to train a multinomial Naive Bayes classifier (Manning et al. , 2008, Ch. 13) which we assign to nb_model. | A képzési adatokat egy multinomiális Naive Bayes-osztályozó (Manning et al. , 2008, 13. fejezet) képzésére használjuk, amelyet az nb_modelhez rendelünk. | A tanulási adatokat egy multinomiális Naive Bayes-osztályozó (Manning et al. , 2008, 13. fejezet) képzésére használjuk, amelyet az nb_modelhez rendelünk. | A tanulási adatokat egy multinomiális naív Bayes módszeren alapuló klasszifikációs algoritmus (Manning et al., 2008,13. fejezet) tanítására használjuk, amelyet az nb_modelhez rendelünk. | Lexikai cserét hajtottunk végre |
| To test how well this model predicts whether an Inaugural Address predates the war, we predict the code for the test data, and make a table in which the rows show the prediction and the columns show the actual value of the is_prewar variable. | Annak teszteléséhez, hogy ez a modell mennyire jól jósolja meg, hogy egy beiktaási beszéd a háború előtt történt-e, megjósoljuk a kódot a tesztadatokra, és készítünk egy táblázatot, amelyben a sorok az előrejelzést, az oszlopok pedig az is_prewar változó tényleges értékét mutatják. | Annak teszteléséhez, hogy ez a modell mennyire jól jelzi előre, hogy egy beiktatási beszéd a háború előtt történt-e, előre jelezzük a kódot a tesztadatokra, és készítünk egy táblázatot, amelyben a sorok az előrejelzést, az oszlopok pedig az is_prewar változó tényleges értékét mutatják. | A modell klasszifikációs képességének vizsgálatára, azaz annak eldöntésére, hogy egy beiktatásib beszéd a háború előtt vagy után hangzott-e el, az alogrimtussal elkésszíttetjük a tesztadatokra alapozott előrejelzést, majd készítünk egy táblázatot, amelyben a sorok az előrejelzést, az oszlopok pedig az is_prewar változó tényleges értékét mutatják. | Szórendi és lexikai csere |

A tanulmány további részének gépi fordítása, részleges és teljes utószerkeszétssel

| Forrásnyelvi szegmens | DeepL gépi fordítás | Részleges utószerkesztés | Teljes utószerkesztés |
|---|---|---|---|
| In unsupervised machine learning approaches, no coding rules are specified and no annotated training data is provided. | A felügyelet nélküli gépi tanulási megközelítéseknél nincsenek kódolási szabályok meghatározva, és nincsenek megjegyzésekkel ellátott képzési adatok. | A felügyelet nélküli gépi tanulási módszereknél nincsenek kódolási szabályok meghatározva, és nincsenek megjegyzésekkel ellátott tanulási adatok sem. | A felügyelet nélküli gépi tanulásnál nincsenek sem előre meghatározott kódolási szabályok, **sem további információkat tartalmazó adatok a betanításhoz.** |
| Instead, an algorithm comes up with a model by identifying certain patterns in text. | Ehelyett egy algoritmus a szövegben található bizonyos minták azonosításával alkot egy modellt. | Ehelyett egy algoritmus a szövegben található bizonyos minták azonosításával alkot egy modellt. | Ehelyett egy algoritmus a szövegben található mintázatok azonosításával **hoz létre** modellt. |
| The only influence of the researcher is the specification of certain parameters, such as the number of categories into which documents are classif ied. | A kutató egyetlen befolyása bizonyos paraméterek megadása, például a kategóriák száma, amelyekbe a dokumentumokat besorolja. | A kutató egyetlen befolyásolási lehetősége bizonyos paraméterek megadása, például a kategóriák száma, amelyekbe a dokumentumokat besorolja. | A kutató egyetlen befolyásolási lehetősége bizonyos paraméterek megadása, **például a képzendő csoportok száma.** |
| Popular examples are topic modeling for automatically classifying documents based on an underlying topical structure (Blei et al., 2003; Roberts et al., 2014) and the “Wordfish” parametric factor model (Proksch & Slapin, 2009) for scaling documents on a single underlying dimension, such as left-right ideology. | Népszerű példák a témamodellezés a dokumentumok automatikus osztályozására egy mögöttes topikstruktúra alapján (Blei et al., 2003; Roberts et al., 2014) és a "Wordfish" parametrikus faktor modell (Proksch & Slapin, 2009) a dokumentumok egyetlen mögöttes dimenzió, például a bal-jobb ideológia alapján történő skálázására. | **Széleskörűen** alkalmazott példája az ilyen algoritmusok használatának a témamodellezés, amikor dokumentumokat automatikus osztályozunk egy, még nem ismert témastruktúra alapján (Blei et al., 2003; Roberts et al., 2014)-, és a "Wordfish" parametrikus faktor modell (Proksch & Slapin, 2009) a dokumentumok egyetlen dimenzió, például a bal-jobb ideológia alapján történő skálázására. | Az ilyen algoritmusok használatának **széleskörűen** alkalmazott példája a témamodellezés, amikor dokumentumokat automatikus osztályozunk egy, még nem ismert témastruktúra alapján (Blei et al., 2003; Roberts et al. , 2014)-, és a "Wordfish" parametrikus faktor modellt (Proksch & Slapin, 2009) **használjuk a** dokumentumok egyetlen dimenzió, például a bal-jobb ideológia alapján történő skálázására. |

| | | | |
|---|---|---|---|
| Grimmer and Stewart (2013) argue that supervised and unsupervised machine learning are not competitor methods, but fulfill different purposes and can very well be used to complement each other. | Grimmer és Stewart (2013) szerint a felügyelt és a felügyelet nélküli gépi tanulás nem konkurens módszerek, hanem különböző célokat töltenek be, és nagyon is jól kiegészíthetik egymást. | Grimmer és Stewart (2013) szerint a felügyelt és a felügyelet nélküli gépi tanulás nem konkurens módszerek, hanem különböző célokat töltenek be, és nagyon is jól kiegészíthetik egymást. | Grimmer és Stewart (2013) szerint a felügyelt és a felügyelet nélküli gépi tanulás nem konkurens módszerek, hanem különböző feladatokat látnak el, és nagyon is jól kiegészíthetik egymást. |
| Supervised methods are the **most suitable** approach if documents need to be placed in predetermined categories, because it is unlikely that an unsupervised method will yield a categorization that **reflects** these categories and how the researcher interprets them. | A felügyelt módszerek a legmegfelelőbb megközelítés, ha a dokumentumokat előre meghatározott kategóriákba kell sorolni, mert nem valószínű, hogy egy felügyelet nélküli módszer olyan kategorizálást fog eredményezni, amely tükrözi ezeket a kategóriákat és azt, hogy a kutató hogyan értelmezi azokat. | A felügyelt módszerek a **legmegfelelőbbek** akkor, ha a dokumentumokat előre meghatározott kategóriákba kell sorolni, mert nem valószínű, hogy egy felügyelet nélküli módszer olyan kategorizálást fog eredményezni, amely **tükrözi** ezeket a kategóriákat és azt, hogy a kutató hogyan értelmezi azokat. | A felügyelt módszerek alkalmazása akkor a **legcélszerűbb,** ha a dokumentumokat előre meghatározott kategóriákba kell sorolni, mert nem valószínű, hogy egy felügyelet nélküli módszer olyan kategorizálást fog eredményezni, amely **megfelel** ezeknek a kategóriáknak és lehetővé teszi a kategorizálás kutatói értelmezését. |
| The advantage of the somewhat unpredictable nature of unsupervised methods is that it can come up with categories that the researchers had not considered. | A nem felügyelt módszerek némileg kiszámíthatatlan természetének előnye, hogy olyan kategóriákkal is előállhat, amelyekre a kutatók nem gondoltak. | A nem felügyelt módszerek némileg kiszámíthatatlan természetének előnye, hogy olyan kategóriák is létrejöhetnek, amelyekre a kutatók nem gondoltak. | A nem felügyelt módszerek némileg kiszámíthatatlan természetének előnye, hogy olyan kategóriák is létrejöhetnek, amelyekre a kutatók nem gondoltak. |
| (Conversely, this may also present challenges for post-hoc interpretation when results are unclear.) | (Ezzel szemben ez kihívást jelenthet az utólagos értelmezés számára is, amikor az eredmények nem egyértelműek.) | (**Ezzel szemben ez kihívást jelenthet** az utólagos értelmezés számára is, amikor az eredmények nem egyértelműek.) | (Amikor az eredmények nem egyértelműek, **ez kihívást jelenthet az** utólagos értelmezés során.) |
| To demonstrate the essence of unsupervised learning, the example below shows how to fit a topic model in R using the topicmodels package (Grun & Hornik, 2011). | A felügyelet nélküli tanulás lényegének bemutatására az alábbi példa azt mutatja be, hogyan lehet egy témamodellt R-ben a topicmodels csomag (Grun & Hornik, 2011) segítségével illeszteni. | A felügyelet nélküli tanulás lényegének bemutatására az alábbi példa szolgál, mely azt **mutatja be,** hogyan lehet egy témamodellt R-ben a topicmodels csomag (Grun & Hornik, 2011) segítségével **illeszteni.** | A felügyelet nélküli tanulás lényegének bemutatására az alábbi példa szolgál, mely azt **demonstrálja,** hogyan lehet **illeszteni** egy témamodellt az R-ben a topicmodels program (Grun & Hornik, 2011) segítségével. |

| | | | |
|---|---|---|---|
| To focus more specifically on topics within the inaugural addresses, and to increase the number of texts to model, we first split the texts by paragraph and create a new vDTM. | Ahhoz, hogy pontosabban a beiktatási beszédeken belüli témákra összpontosítsunk, és hogy növeljük a modellezendő szövegek számát, először bekezdésenként felosztjuk a szövegeket, és létrehozunk egy új DTM-et. | Ahhoz, hogy pontosabban a összpontosítsunk a beiktatási beszédekben felmerülő témákra, és hogy növeljük a modellezendő szövegek számát, először bekezdésenként felosztjuk a szövegeket, és létrehozunk egy új DTM-et. | Ahhoz, hogy pontosabban összpontosíthassunk a beiktatási beszédekben felmerülő témákra, és hogy növeljük a modellezendő szövegek számát, először bekezdésenként felosztjuk a szövegeket, és létrehozunk egy új DTM-et. |
| From this DTM we remove terms with a document frequency of five and lower to reduce the size of the vocabulary (less important for current example) and use quanteda's convert function to convert the DTM to the format used by topicmodels. | Ebből a DTM-ből eltávolítjuk az ötös vagy annál kisebb dokumentumfrekvenciájú kifejezéseket, hogy csökkentsük a szókincs méretét (ami a jelenlegi példa szempontjából kevésbé fontos), és a quanteda convert funkciójával a DTM-et a topicmodels által használt formátumba konvertáljuk. | Ebből a DTM-ből eltávolítjuk az öt, vagy annál kisebb gyakorisáú kifejezéseket, hogy csökkentsük a szavak számát (ez a jelenlegi példa szempontjából kevésbé fontos), és a quanteda convert funkciójával a DTM-et a topicmodels által használt formátumba konvertáljuk. | Ebből a DTM-ből eltávolítjuk az öt, vagy annál kisebb gyakoriságú kifejezéseket, hogy csökkentsük a szavak számát (ez a jelenlegi példa szempontjából kevésbé fontos), és a quanteda convert funkciójával a DTM-et a topicmodels által használt formátumba konvertáljuk. |
| We then train a vanilla LDA topic model(Blei et al. , 2003) with five topics—using a fixed seed to make the results reproducible, since LDA is non-deterministic. | Ezután egy vanília LDA témamodellt (Blei et al. , 2003) képzünk ki öt témával - fix magot használunk, hogy az eredmények reprodukálhatók legyenek, mivel **az LDA nem determinisztikus**. | Ezután egy vanílla LDA témamodellt (Blei et al. , 2003) tanítunk be öt témával – **az LDA nem determinisztikus**, ezért az eredmények reprodukálhatósága érdekében rögzített kiindulási szimulációs értékeket használunk. | Ezután egy vanílla LDA témamodellt (Blei et al. , 2003) tanítunk be öt témával. **Az LDA nem determinisztikus**, ezért az eredmények reprodukálhatósága érdekében rögzített kiindulási szimulációs értékeket használunk. |
| Various statistics can be used to describe, explore and analyze a text corpus. | A szövegkorpuszok leírására, feltárására és elemzésére különböző statisztikák használhatók. | A szövegkorpuszok leírására, feltárására és elemzésére különböző statisztikák használhatók. | A szövegkorpuszok leírására, feltárására és elemzésére különböző statisztikák használhatók. |
| An example of a **popular** technique is to rank the information value of words inside a corpus and then visualize the most informative words as a word cloud to get a quick indication of what a corpus is about. | Egy népszerű technika például a korpuszban található szavak információértékének rangsorolása, majd a leginformatívabb szavak szófelhő formájában történő megjelenítése, | Egy n**épszerű** eljárás például a korpuszban található szavak információértékének rangsorolása, majd a leginformatívabb szavak szófelhő formájában történő megjelenítése, | **Gyakran alkalmazott** eljárás például a korpuszban található szavak információértékének rangsorolása, majd a leginformatívabb szavak szófelhő formájában történő megjelenítése **annak érdekében, hogy** gyors |

| | | | |
|---|---|---|---|
| | hogy gyors képet kapjunk arról, miről szól a korpusz. | hogy gyors képet kapjunk arról, miről szól a korpusz. | képet kapjunk a korpusz tartalmáról. |
| Text statistics (e. g. , average word and sentence length, word and syllable counts) are also commonly used as an operationalization of concepts such as readability (Flesch, 1948) or lexical diversity (McCarthy & Jarvis, 2010). | A szövegstatisztikákat (pl. átlagos szó- és mondathossz, szó- és szótagszám) szintén gyakran használják olyan fogalmak operacionalizálására, mint az olvashatóság (Flesch, 1948) vagy a lexikai sokszínűség (McCarthy & Jarvis, 2010). | A szövegstatisztikákat (pl. átlagos szó- és mondathossz, szó- és szótagszám) szintén gyakran használják olyan fogalmak operacionalizálására, mint az olvashatóság (Flesch, 1948) vagy a lexikai sokszínűség (McCarthy & Jarvis, 2010). | A szövegstatisztikákat (pl. átlagos szó- és mondathossz, szó- és szótagszám) szintén gyakran használják az olyan fogalmak operacionalizálására, mint **például** az olvashatóság (Flesch, 1948) vagy a lexikai sokszínűség (McCarthy & Jarvis, 2010). |
| A wide range of such measures is available in R with the koRpus package (Michalke, 2017). | Az ilyen mérőszámok széles skálája elérhető az R-ben a koRpus csomaggal (Michalke, 2017). | Az ilyen mérőszámok széles skálája elérhető az R-ben a koRpus csomaggal (Michalke, 2017). | Az ilyen mérőszámok széles skálája elérhető az R-ben a koRpus csomaggal (Michalke, 2017). |
| Furthermore, there are many **useful applications** of calculating term and document similarities (which are often based on the inner product of a DTM or transposed DTM), such as analyzing semantic relations between words or concepts and measuring content homogeneity. | Továbbá a terminus- és dokumentumhasonlóságok kiszámításának (amelyek gyakran egy DTM vagy transzponált DTM belső szorzatán alapulnak) számos hasznos alkalmazása van, például a szavak vagy fogalmak közötti szemantikai kapcsolatok elemzése és a tartalmi homogenitás mérése. | A terminus- és dokumentumhasonlóságok kiszámításának (amelyek gyakran egy DTM vagy transzponált DTM belső szorzatán alapulnak) számos **hasznos alkalmazása** van, például a szavak vagy fogalmak közötti szemantikai kapcsolatok elemzése és a tartalmi homogenitás mérése. | A terminus- és dokumentumhasonlóságok kiszámításának (amelyek gyakran egy DTM vagy transzponált DTM belső szorzatán alapulnak) számos **felhasználási lehetősége** van, például a szavak vagy fogalmak közötti szemantikai kapcsolatok elemzése és a tartalmi homogenitás mérése. |
| Both techniques are supported in quanteda, corpustools (Welbers & Van Atteveldt, 2016), or dedicated packages such as textreuse (Mullen, 2016a) for text overlap. | Mindkét technikát támogatja a quanteda, a corpustools (Welbers & Van Atteveldt, 2016) vagy a szövegek átfedésére szolgáló dedikált csomagok, például a textreuse (Mullen, 2016a). | Mindkét technikát támogatja a quanteda, a corpustools (Welbers & Van Atteveldt, 2016) vagy a szövegek átfedésének vizsgálatára szolgáló speciális csomagok, például a textreuse (Mullen, 2016a). | Mindkét technikát támogatja a quanteda, és a corpustools (Welbers & Van Atteveldt, 2016) **továbbá** a szövegek átfedésének **vizsgálatára** szolgáló speciális csomagok, például a textreuse (Mullen, 2016a). |
| A particularly useful technique is to compare the term frequencies of two corpora, or between two subsets of the same corpus. | Különösen hasznos technika két korpusz vagy ugyanazon korpusz két részhalmazának kifejezésfrekvenciáit összehasonlítani. | Különösen hasznos, **ha** két korpuszban, vagy ugyanazon korpusz két részhalmazában hasonlítjuk össze az egyes terminusok gyakoriságát. | Különösen hasznos**, ha** két korpuszban, vagy ugyanazon korpusz két részében hasonlítjuk össze az egyes terminusok gyakoriságát. |

| | | | |
|---|---|---|---|
| For instance, to see which words are **more likely** to occur in documents about a certain topic. | Például, hogy megnézzük, mely szavak fordulnak elő nagyobb **valószínűséggel** egy adott témájú dokumentumokban. | Ilyen például, ha megnézzük, mely szavak fordulnak elő nagyobb **valószínűséggel** egy adott témájú dokumentumban. | **Ilyenkor például megvizsgáljuk,** hogy mely szavak fordulnak elő nagyobb **gyakorisággal** egy adott témájú dokumentumban. |
| In addition to providing a way to quickly explore how this topic is discussed in the corpus, this can provide input for developing **better** queries. | Ez amellett, hogy lehetőséget nyújt annak gyors feltárására, hogy az adott témát hogyan tárgyalják a korpuszban, inputot adhat **jobb** lekérdezések kidolgozásához. | Ez amellett, hogy lehetőséget nyújt annak gyors feltárására, hogy az adott témát hogyan tárgyalják a korpuszban, inputot adhat **jobb** lekérdezések kidolgozásához. | Ez amellett, hogy lehetőséget nyújt annak gyors feltárására, hogy az adott témát hogyan tárgyalják a korpuszban, ötleteket adhat **hatékonyabb** lekérdezések kidolgozásához. |
| In the following example we show how to perform this technique in quanteda **(results in Figure 2).** | A következő példában bemutatjuk, hogyan lehet ezt a technikát a quanteda-ban végrehajtani (**eredmények a 2. ábrán).** | A következő példában bemutatjuk, hogyan lehet ezt a technikát a quanteda-ban végrehajtani **(eredmények a 2. ábrán).** | A következő példában bemutatjuk, hogyan lehet ezt az eljárást a quanteda-ban végrehajtani **(Eredményeinket a 2. ábra szemlélteti).** |
| Itt a szignált χ2 asszociációs mérőszám azt mutatja, hogy az "america", "american" és "first" szavakat Trump sokkal gyakrabban használta, mint Obama, míg a "us", "can", "freedom", "peace" és "liberty" szavakat Obama sokkal gyakrabban használta, mint Trump. | Itt a szignált χ2 asszociációs mérőszám azt mutatja, hogy az "america", "american" és "first" szavakat Trump sokkal gyakrabban használta, mint Obama, míg a "us", "can", "freedom", "peace" és "liberty" szavakat Obama sokkal gyakrabban használta, mint Trump. | Itt a χ2-val jeölt asszociációs mérőszám azt mutatja, hogy az "america", "american" és "first" szavakat Trump sokkal gyakrabban használta, mint Obama, míg az "us", "can", "freedom", "peace" és "liberty" szavakat Obama használta sokkal gyakrabban, mint Trump. | Itt a χ2-val jeölt asszociációs mérőszám azt mutatja, hogy az "america", "american" és "first" szavakat Trump sokkal gyakrabban használta, mint Obama, míg az "us", "can", "freedom", "peace" és "liberty" szavakat Obama használta sokkal gyakrabban, mint Trump. |
| The data preparation and bag-of-words analysis techniques discussed above are the basis for the majority of the text analysis approaches that are currently used in communication research. | A fentiekben tárgyalt adatelőkészítési és szóhalmazelemzési technikák képezik az alapját a kommunikációkutatásban jelenleg alkalmazott szövegelemzési megközelítések többségének. | A fentiekben tárgyalt adatelőkészítési és szóhalmazelemzési technikák képezik **az alapját** a kommunikációkutatásban jelenleg alkalmazott szövegelemzési módszerek többségének. | A fentiekben tárgyalt adatelőkészítési és szóhalmazelemzési módszerek képezik a kommunikációkutatásban jelenleg alkalmazott a szövegelemzési módszerek többségének **az alapját.** |
| For certain types of analyses, however, techniques might be | Bizonyos típusú elemzésekhez azonban szükség lehet olyan | Bizonyos típusú elemzésekhez azonban szükség lehet olyan | Bizonyos típusú elemzésekhez azonban szükség lehet olyan |

| | | | |
|---|---|---|---|
| required that rely on external software modules, are more computationally demanding, or that are more complicated to use. | technikákra, amelyek külső szoftvermodulokra támaszkodnak, számításigényesebbek vagy bonyolultabbak a használatukban. | technikákra, amelyek külső szoftvermodulokra támaszkodnak, számításigényesebbek vagy használatuk bonyolultab. | technikákra, amelyek külső szoftvermodulokra támaszkodnak, számításigényesebbek vagy használatuk bonyolultab. |
| In this section we briefly elaborate on some of these advanced approaches that are worth taking note of. | Ebben a szakaszban röviden bemutatunk néhány ilyen fejlett megközelítést, amelyeket érdemes figyelembe venni. | Ebben a fejezetben röviden bemutatunk néhány ilyen fejlett módszert, amelyeket érdemes figyelembe venni. | Ebben a fejezetben röviden bemutatunk néhány ilyen fejlett módszert, amelyeket érdemes figyelembe venni. |
| In addition to the preprocessing **techniques** discussed in the data preparation section, there are powerful preprocessing techniques that rely on more advanced natural language processing (NLP). | Az adatelőkészítési szakaszban tárgyalt előfeldolgozási **technikákon** kívül vannak olyan hatékony előfeldolgozási technikák, amelyek a fejlettebb természetes nyelvi feldolgozásra (NLP) támaszkodnak. | Az adatelőkészítéssel foglalkozó fejezetben tárgyalt előfeldolgozási **technikákon** kívül vannak olyan hatékony előfeldolgozási technikák is, amelyek a fejlettebb természetes nyelvi feldolgozásra (NLP) támaszkodnak. | Az adatelőkészítéssel foglalkozó fejezetben tárgyalt előfeldolgozási **módszereken** kívül vannak olyan hatékony előfeldolgozási eljárások is, amelyek a fejlettebb, természetes nyelvi feldolgozásra (NLP) támaszkodnak. |
| At present, these **techniques** are not available in native R, but rely on external software modules that often have to be installed outside of R. | Jelenleg ezek a **technikák** nem állnak rendelkezésre az R-ben, hanem külső szoftvermodulokra támaszkodnak, amelyeket gyakran az R-en kívül kell telepíteni. | Jelenleg ezek **a technikák** nem állnak rendelkezésre az R-ben, hanem külső szoftvermodulokra támaszkodnak, amelyeket gyakran az R-en kívül kell telepíteni. | Jelenleg ezek a **módszerek** nem állnak rendelkezésre az R-ben, hanem külső szoftvermodulokra támaszkodnak, amelyeket az R-en kívül kell telepíteni. |
| Many of the advanced NLP techniques are also much more computationally demanding, thus taking more time **to perform.** | A fejlett NLP-technikák közül sok számításigényesebb is, így több időt vesz igénybe a **végrehajtásuk.** | A fejlett NLP-technikák közül sok számításigényesebb is, így több időt vesz igénybe a **végrehajtásuk.** | A fejlett NLP-technikák közül sok számításigényesebb is, így a **végrehajtásuk** több időt vesz igénybe. |
| Another complication is that these techniques are language specific, and often only available for English and a few other big languages. | További bonyodalom, hogy ezek a technikák nyelvspecifikusak, és gyakran csak az angol és néhány más nagy nyelvhez állnak rendelkezésre. | További nehézség, hogy ezek a technikák nyelvspecifikusak, és gyakran csak az angol és néhány más nagy nyelvhez állnak rendelkezésre. | További nehézség, hogy ezek az eljárások nyelvspecifikusak, és gyakran csak az angol és néhány más nagy nyelvhez kapcsolódnak. |
| Several R packages provide interfaces for external NLP modules, so that once these | Számos R csomag biztosít interfészeket külső NLP modulokhoz, így ha ezeket a modulokat egyszer már | Néhány R csomag biztosít interfészeket külső NLP modulokhoz, így ha ezeket a modulokat egyszer már | |

| | | | |
|---|---|---|---|
| modules have been installed, they can easily be used from within R. | telepítettük, könnyen használhatjuk őket R-ből. | telepítettük, könnyen használhatjuk őket R-ből. | |
| The coreNLP package (Arnold & Tilton, 2016) provides bindings for Stanford CoreNLP java library (Manning et al. | A coreNLP csomag (Arnold & Tilton, 2016) kötéseket biztosít a Stanford CoreNLP java könyvtárhoz (Manning et al. | A coreNLP csomag (Arnold & Tilton, 2016) kapcsolatot biztosít a Stanford CoreNLP java könyvtárhoz (Manning et al. | |
| , 2014), which is a full NLP parser for English, that also supports (albeit with limitations) Arabic, Chinese, French, German, and Spanish. | , 2014), amely egy teljes körű NLP elemző az angol nyelvhez, amely (bár korlátozásokkal) támogatja az arab, kínai, francia, német és spanyol nyelveket is. | , 2014), amely egy teljes körű NLP elemző az angol nyelvhez, amely (bár korlátozásokkal) támogatja az arab, kínai, francia, német és spanyol nyelveket is. | |
| The spacyr package (Benoit & Matsuo, 2017) provides an interface for the spaCy module for Python, which is comparable to CoreNLP but is faster, and supports English and (again with some limitations) German and French. | A spacyr csomag (Benoit & Matsuo, 2017) interfészt biztosít a pythoni spaCy modulhoz, amely a CoreNLP-hez hasonló, de gyorsabb, és támogatja az angol nyelvet, valamint (szintén bizonyos korlátozásokkal) a németet és a franciát. | A spacyr szoftver (Benoit & Matsuo, 2017) interfészt biztosít a python spaCy moduljához, amely a CoreNLP-hez hasonló, de gyorsabb, és támogatja az angol nyelvet, valamint (szintén bizonyos korlátozásokkal) a németet és a franciát is. | |
| A third package, cleanNLP (Arnold, 2017a), conveniently wraps both CoreNLP and spaCy, and also includes a minimal back-end that does not rely on external dependencies. | Egy harmadik csomag, a cleanNLP (Arnold, 2017a) kényelmesen csomagolja mind a CoreNLP-t, mind a spaCy-t, és tartalmaz egy minimális back-endet is, amely nem támaszkodik külső függőségekre. | Egy harmadik szoftver, a cleanNLP (Arnold, 2017a) kényelmesen csomagolja mind a CoreNLP-t, mind a spaCy-t, és tartalmaz egy minimális back-endet is, amely nem támaszkodik külső függőségekre. | |
| This way it can be used as a swiss army knife, choosing the approach that best suits the occasion and for which the back-end is available, but with standardized output and methods. | Így svájci bicskaként használható, az alkalomhoz legjobban illő megközelítést választva, amelyhez a back-end rendelkezésre áll, de egységesített kimenettel és módszerekkel. | Így svájci bicskaként használható, az alkalomhoz legjobban illő megközelítést választva, amelyhez a back-end rendelkezésre áll, de egységesített kimenettel és módszerekkel. | |
| Several R packages provide interfaces for external NLP modules, so that once these | Számos R csomag biztosít interfészeket külső NLP modulokhoz, így ha ezeket a modulokat egyszer már | Néhány R csomag biztosít interfészeket külső NLP modulokhoz, így ha ezeket a modulokat egyszer már | |

| | | | |
|---|---|---|---|
| modules have been installed, they can easily be used from within R. | telepítettük, könnyen használhatjuk őket R-ből. | telepítettük, könnyen használhatjuk őket R-ből. | |
| The coreNLP package (Arnold & Tilton, 2016) provides bindings for Stanford CoreNLP java library (Manning et al. | A coreNLP csomag (Arnold & Tilton, 2016) kötéseket biztosít a Stanford CoreNLP java könyvtárhoz (Manning et al. | A coreNLP csomag (Arnold & Tilton, 2016) kapcsolatot biztosít a Stanford CoreNLP java könyvtárhoz (Manning et al. | A coreNLP program (Arnold & Tilton, 2016) kapcsolatot biztosít a Stanford CoreNLP java könyvtárhoz (Manning et al. |
| , 2014), which is a full NLP parser for English, that also supports (albeit with limitations) Arabic, Chinese, French, German, and Spanish. | , 2014), amely egy teljes körű NLP elemző az angol nyelvhez, amely (bár korlátozásokkal) támogatja az arab, kínai, francia, német és spanyol nyelveket is. | , 2014), amely egy teljes körű NLP elemző az angol nyelvhez, amely (bár korlátozásokkal) támogatja az arab, kínai, francia, német és spanyol nyelveket is. | , 2014), amely egy teljes körű NLP elemző szoftver az angol nyelvhez, és amelyik (korlátozásokkal ugyan) támogatja az arab, kínai, francia, német és spanyol nyelveket is. |
| The spacyr package (Benoit & Matsuo, 2017) provides an interface for the spaCy module for Python, which is comparable to CoreNLP but is faster, and supports English and (again with some limitations) German and French. | A spacyr csomag (Benoit & Matsuo, 2017) interfészt biztosít a pythoni spaCy modulhoz, amely a CoreNLP-hez hasonló, de gyorsabb, és támogatja az angol nyelvet, valamint (szintén bizonyos korlátozásokkal) a németet és a franciát. | A spacyr szoftver (Benoit & Matsuo, 2017) interfészt biztosít a python spaCy moduljához, amely a CoreNLP-hez hasonló, de gyorsabb, és támogatja az angol nyelvet, valamint (szintén bizonyos korlátozásokkal) a németet és a franciát is. | A spacyr szoftver (Benoit & Matsuo, 2017) interfészt biztosít a python spaCy moduljához, amely a CoreNLP-hez hasonló, de gyorsabb, és támogatja az angol nyelvet, valamint (szintén bizonyos korlátozásokkal) a németet és a franciát is. |
| A third package, cleanNLP (Arnold, 2017a), conveniently wraps both CoreNLP and spaCy, and also includes a minimal back-end that does not rely on external dependencies. | Egy harmadik csomag, a cleanNLP (Arnold, 2017a) kényelmesen csomagolja mind a CoreNLP-t, mind a spaCy-t, és tartalmaz egy minimális back-endet is, amely nem támaszkodik külső függőségekre. | Egy harmadik szoftver, a cleanNLP (Arnold, 2017a) kényelmesen csomagolja mind a CoreNLP-t, mind a spaCy-t, és tartalmaz egy minimális back-endet is, amely nem támaszkodik külső függőségekre. | Egy harmadik szoftver, a cleanNLP (Arnold, 2017a) kényelmesen integrálja mind a CoreNLP-t, mind a spaCy-t, és tartalmaz egy minimális háttérszámításokat végző molult is, amely nem támaszkodik külső szoftverekre. |
| This way it can be used as a swiss army knife, choosing the approach that best suits the occasion and for which the back-end is available, but with standardized output and methods. | Így svájci bicskaként használható, az alkalomhoz legjobban illő megközelítést választva, amelyhez a back-end rendelkezésre áll, de egységesített kimenettel és módszerekkel. | Így svájci bicskaként használható, az alkalomhoz legjobban illő megközelítést választva, amelyhez a back-end rendelkezésre áll, de egységesített kimenettel és módszerekkel. | Így egyfajta “svájci bicskaként” használható, a feladathoz legjobban illő módszert választva, amelyhez a háttérszámításokat végző modul egységesített kimenettel és módszerekkel áll rendelkezésre. |

| | | | |
|---|---|---|---|
| Several R packages provide interfaces for external NLP modules, so that once these modules have been installed, they can easily be used from within R. | Számos R csomag biztosít interfészeket külső NLP modulokhoz, így ha ezeket a modulokat egyszer már telepítettük, könnyen használhatjuk őket R-ből. | Néhány R csomag biztosít interfészeket külső NLP modulokhoz, így ha ezeket a modulokat egyszer már telepítettük, könnyen használhatjuk őket R-ből. | |
| Lemmatization. | Lemmatizáció. | Lemmatizáció. | Lemmatizáció. |
| Lemmatization fulfills a similar purpose as stemming, but instead of cutting off the ends of terms to normalize them, a dictionary is used to replace terms with their lemma. | A lemmatizálás hasonló célt szolgál, mint a törzselés, de ahelyett, hogy a kifejezések végét levágnánk, hogy normalizáljuk őket, egy szótárat használunk a kifejezések helyettesítésére a lemma helyett. | A lemmatizálás hasonló célt szolgál, mint a szótőképzés, de ahelyett, hogy a szavak végét levágnánk, hogy normalizáljuk őket, egy szótárat használunk a terminusok helyettesítésére a lemmával. | A lemmatizálás hasonló célt szolgál, mint a szótőképzés, de ahelyett, hogy a szavak végét levágnánk a normalizálásuk érdekében, egy szótárat használunk arra, hogy a terminusokat lemmáikkal helyettesítsük. |
| The main advantage of this approach is that it can more accurately normalize different verb forms—such as “gave” and “give” in the example—which is particularly important for heavily inflected languages such as Dutch or German. | Ennek a megközelítésnek a fő előnye, hogy pontosabban normalizálja a különböző igealakokat - például a példában a "gave" és a "give" -, ami különösen fontos az olyan erősen flektált nyelvek esetében, mint a holland vagy a német. | Ennek a megközelítésnek a fő előnye, hogy pontosabban normalizálja a különböző igealakokat - például a példában a "gave" és a "give" -, ami különösen fontos az olyan erősen ragozó nyelvek esetében, mint a holland vagy a német. | Ennek a módszernek az a fő előnye, hogy pontosabban normalizálja a különböző igealakokat - például a példában a "gave" és a "give" -, ez különösen fontos az olyan erősen ragozó nyelvek esetében, mint a holland vagy a német. |
| Part-of-speech tagging. | Beszédrészek címkézése. | Beszédrészek címkézése. | A beszédrészek címkézése. |
| POS tags are morpho-syntactic categories for words, such as nouns, verbs, articles and adjectives. | A POS-címkék a szavak morfoszintaktikai kategóriái, például főnevek, igék, cikkek és melléknevek. | A POS-címkék a szavak morfoszintaktikai kategóriái, például főnevek, igék, névelők és melléknevek. | A POS-címkék a szavak morfoszintaktikai kategóriái, például főnevek, igék, névelők és melléknevek. |
| In the example we see three proper nouns (PROPN), a verb (VERB) an adjective (ADJ), two nouns (NOUN), and punctuation (PUNCT). | A példában három tulajdonnév (PROPN), egy ige (VERB), egy melléknév (ADJ), két főnév (NOUN) és írásjel (PUNCT) látható. | A példában három tulajdonnév (PROPN), egy ige (VERB), egy melléknév (ADJ), két főnév (NOUN) és írásjel (PUNCT) látható. | A példában három tulajdonnév (PROPN), egy ige (VERB), egy melléknév (ADJ), két főnév (NOUN) és írásjel (PUNCT) látható. |
| This information can be used to focus an analysis on certain types of grammar categories, for | Ez az információ felhasználható arra, hogy az elemzést bizonyos típusú nyelvtani kategóriákra | Ez az információ felhasználható arra, hogy az elemzést bizonyos típusú nyelvtani kategóriákra | Ez az információ felhasználható arra, hogy az elemzést bizonyos típusú nyelvtani kategóriákra |

| | | | |
|---|---|---|---|
| example, using nouns and proper names to measure similar events in news items (Welbers, Van Atteveldt, Kleinnijenhuis, & Ruigrok, 2016), or using adjectives to focus on subjective language (De Smedt & Daelemans, 2012). | összpontosítsuk, például a főnevek és tulajdonnevek segítségével a hírekben szereplő hasonló események mérésére (Welbers, Van Atteveldt, Kleinnijenhuis, & Ruigrok, 2016), vagy a melléknevek segítségével a szubjektív nyelvre összpontosítva (De Smedt & Daelemans, 2012). | összpontosítsuk, például a köznevek és tulajdonnevek segítségével a hírekben szereplő hasonló események mérésére (Welbers, Van Atteveldt, Kleinnijenhuis, & Ruigrok, 2016), vagy a melléknevek segítségével a szubjektív nyelvre összpontosítva (De Smedt & Daelemans, 2012). | összpontosítsuk, például a köznevek és tulajdonnevek segítségével a hírekben szereplő hasonló események számszerűsítésére (Welbers, Van Atteveldt, Kleinnijenhuis, & Ruigrok, 2016), vagy a melléknevek segítségével a szubjektív elemeket tartalmazó nyelvhasználatra összpontosítva (De Smedt & Daelemans, 2012). |
| Similarly, it is a good approach for filtering out certain types of words, such as articles or pronouns. | Hasonlóképpen jó megközelítés bizonyos szófajok, például cikkek vagy névmások kiszűrésére. | Hasonlóképpen jó megközelítés bizonyos szófajok, például névelők vagy névmások kiszűrésére. | Hasonlóképpen jó megközelítés bizonyos szófajok, például névelők vagy névmások kiszűrésére. |
| Named entity recognition. | Megnevezett entitások felismerése. | Megnevezett entitások felismerése. | Megnevezett entitások felismerése. |
| Named entity recognition is a technique for identifying whether a word or sequence of words represents an entity and what type of entity, such as a person or organization. | A megnevezett entitások felismerése egy olyan technika, amellyel azonosítható, hogy egy szó vagy szósorozat egy entitást képvisel-e, és milyen típusú entitást, például személyt vagy szervezetet. | A megnevezett entitások felismerése egy olyan módszer, amellyel azonosítható, hogy egy szó vagy szósorozat egy entitást képvisel-e, és ha igen, akkor az milyen, például személyt vagy szervezetet. | A megnevezett entitások felismerése egy olyan módszer, amellyel azonosítható, hogy egy szó vagy szósorozat egy entitást képvisel-e, és ha igen, akkor az milyen, például személyt vagy szervezetet. |
| Both “Bob Smith” and “Alice” are recognized as persons. | A "Bob Smith" és az "Alice" egyaránt személynek minősül. | A "Bob Smith" és az "Alice" egyaránt személynek minősül. | A "Bob Smith" és az "Alice" egyaránt személynek minősül. |
| Ideally, named entity recognition is paired with co-reference resolution. | Ideális esetben a megnevezett entitások felismerése társhivatkozások feloldásával párosul. | Ideális esetben a megnevezett entitások felismerése társhivatkozások feloldásával párosul. | Ideális esetben a megnevezett entitások azonosítása a rájuk történő hivatkozások egységesítésével párosul, és az ugyanarra az entitásra történő különböző hivatkozásokat, például anaforákat (pl. ő, nő, az elnök) csoportosíthatjuk. |
| This is a technique for grouping different references to the same | Ez egy olyan technika, amellyel az ugyanarra az entitásra | Ez egy olyan technika, amellyel az ugyanarra az entitásra | A példamondatban az "ő" szó "Bob Smith"-re utal. |

| | | | |
|---|---|---|---|
| entity, such as anaphora (e.g. he, she, the president). | vonatkozó különböző hivatkozásokat, például anaforákat (pl. ő, nő, az elnök) csoportosíthatjuk. | vonatkozó különböző hivatkozásokat, például anaforákat (pl. ő, nő, az elnök) csoportosíthatjuk. | |
| In the example sentence, the word “his” refers to “Bob Smith”. | A példamondatban az "ő" szó "Bob Smith"-re utal. | A példamondatban az "ő" szó "Bob Smith"-re utal. | Az ilyen hivatkozások azonosítását jelenleg csak a Stanford CoreNLP támogatja, de a spaCy GitHub oldalán folyó vita arra utal, hogy ez a funkció napirenden van. |
| Dependency parsing. | Függőségi elemzés. | Kapcsolati elemzés. | Kapcsolati elemzés. |
| Dependency parsing provides the syntactic relations between tokens, which can be used to analyze texts at the level of syntactic clauses (Van Atteveldt, 2008). | A függőségi elemzés megadja a tokenek közötti szintaktikai kapcsolatokat, amelyek segítségével a szövegek szintaktikai tagmondatok szintjén elemezhetők (Van Atteveldt, 2008). | A kapcsolati elemzés megadja a tokenek közötti szintaktikai kapcsolatokat, amelyek segítségével a szövegek szintaktikai egységek szintjén elemezhetők (Van Atteveldt, 2008). | A kapcsolati elemzés megadja a tokenek közötti szintaktikai kapcsolatokat, amelyek segítségével a szövegek szintaktikai egységek szintjén elemezhetők (Van Atteveldt, 2008). |
| In the spacyr output this information is given in the head_token_i and dep_rel columns, where the former indicates to what token a token is related and the latter indicates the type of relation. | A spacyr kimenetén ez az információ a head_token_i és a dep_rel oszlopokban szerepel, ahol az előbbi azt jelzi, hogy egy token milyen tokenhez kapcsolódik, az utóbbi pedig a kapcsolat típusát. | A spacyr outputjában ez az információ a head_token_i és a dep_rel oszlopokban szerepel, ahol az előbbi azt jelzi, hogy egy token milyen tokenhez kapcsolódik, az utóbbi pedig a kapcsolat típusát adja meg. | A spacyr outputjában ez az információ a head_token_i és a dep_rel oszlopokban szerepel, ahol az előbbi azt jelzi, hogy egy token milyen tokenhez kapcsolódik, az utóbbi pedig a kapcsolat típusát jelöli. |
| For example, we see that “Bob” is related to “Smith” (head_token_i 2) as a compound, thus recognizing “Bob Smith” as a single entity. | Például azt látjuk, hogy a "Bob" a "Smith"-hez (head_token_i 2) összetett viszonyban áll, így a "Bob Smith"-t egyetlen entitásként ismeri fel. | Például azt látjuk, hogy a "Bob" a "Smith"-tel (head_token_i 2) egy egységet alkot, így a "Bob Smith"-t egyetlen entitásként ismeri fel. | Azt látjuk például, hogy a "Bob" a "Smith"-tel (head_token_i 2) egy egységet alkot, így a szoftver a "Bob Smith"-t egyetlen entitásként ismeri fel. |
| Also, since “Smith” is the nominal subject (nsubj) of the verb “gave”, and Alice is the dative case (dative) we know that “Bob Smith” is the one who gives to “Alice”. | Továbbá, mivel "Smith" a "gave" ige nominális alanya (nsubj), Alice pedig a dativus (dative), tudjuk, hogy "Bob Smith" az, aki ad "Alice"-nek. | Továbbá, mivel "Smith" a "gave" ige névszói alanya (nsubj), Alice pedig a dativus, tudjuk, hogy "Bob Smith" az, aki ad "Alice"-nek. | Továbbá, mivel "Smith" a "gave" ige névszói alanya (nsubj), Alice pedig a dativus, (részeshatározó) tudjuk, hogy "Bob Smith" az, aki ad "Alice"-nek. |

| | | | |
|---|---|---|---|
| This type of information can for instance be used to analyze who is attacking whom in news coverage about the Gaza war (Van Atteveldt et al. , 2017). | Az ilyen típusú információk például felhasználhatók annak elemzésére, hogy a gázai háborúról szóló híradásokban ki kit támad meg (Van Atteveldt et al. , 2017). | Az ilyen típusú információk például felhasználhatók annak elemzésére, hogy a gázai háborúról szóló híradásokban ki kit támad meg (Van Atteveldt et al. , 2017). | Az ilyen típusú információk például felhasználhatók annak elemzésére, hogy ki kit támad meg a gázai háborúról szóló híradások alapján (Van Atteveldt et al. , 2017). |
| **Word positions and syntax** | Szóhelyzetek és szintaxis | A szavak helyzete és a szintaxis elemzése | A szavak helyzete és a szintaxis elemzése |
| As discussed above, the bag-of-words representation of texts is memory-efficient and convenient for various types of analyses, and this often outweighs the disadvantage of losing information by dropping the word positions. | Amint azt fentebb tárgyaltuk, a szövegek szózsákos reprezentációja memóriahatékony és kényelmes a különböző típusú elemzésekhez, és ez gyakran ellensúlyozza azt a hátrányt, hogy a szóhelyzetek elhagyásával információveszteséget okoz. | Amint azt fentebb tárgyaltuk, a szövegek szózsákos reprezentációja memóriahatékony és kényelmes a különböző típusú elemzésekhez, és ez gyakran ellensúlyozza azt a hátrányt, hogy az egyes szavak mondaton belüli elhelyezkedésének figyelm kívül hagyásával információveszteséget okoz. | Amint azt korábban bemutattuk, a szövegek szózsákos reprezentációja memóriahatékony és kényelmes lehetőségeket kínál a különböző elemzésekhez. |
| For some analyses, however, the order of words and syntactical properties can be highly beneficial if not crucial. | Bizonyos elemzéseknél azonban a szavak sorrendje és a szintaktikai tulajdonságok igen hasznosak, ha nem döntő fontosságúak lehetnek. | Bizonyos elemzéseknél azonban a szavak sorrendje és a szintaktikai tulajdonságok ismerete igen hasznosak, sőt döntő fontosságúak lehetnek. | Ez általában ellensúlyozza azt a hátrányt, hogy az egyes szavak mondaton belüli elhelyezkedésének figyelmen kívül hagyásával információveszteség keletkezik. |
| In this section we address some text representations and analysis techniques where word positions are maintained. | Ebben a fejezetben néhány olyan szövegreprezentációval és elemzési technikával foglalkozunk, ahol a szóhelyzetek megmaradnak. | Ebben a fejezetben néhány olyan szövegreprezentációval és elemzési technikával foglalkozunk, ahol a szóhelyzetek megmaradnak. | Bizonyos elemzéseknél azonban a szavak sorrendje és a szintaktikai tulajdonságok ismerete igen hasznosak, sőt döntő fontosságúak lehetnek. |
| A simple but potentially powerful solution is to use higher order n-grams. | Egy egyszerű, de potenciálisan hatékony megoldás a magasabb rendű n-grammok használata. | Egy egyszerű, de hatékony megoldás lehet a magasabb rendű n-grammok használata. | Egyszerű, de hatékony megoldás lehet a magasabb rendű n-grammok használata. |
| That is, instead of tokenizing texts into single words (n = 1; unigrams), sequences of two words (n = 2; bigrams), three | Ez azt jelenti, hogy ahelyett, hogy a szövegeket egyetlen szóra (n = 1; unigramok) tokenizálnánk, két szóból (n = 2; bigramok), | Ez azt jelenti, hogy ahelyett, hogy a szövegeket szavanként (n = 1; unigramok) tokenizálnánk, két szóból (n = 2; bigramok), | Ez azt jelenti, hogy ahelyett, hogy a szövegeket szavanként (n = 1; unigramok) tokenizálnánk, két szóból (n = 2; bigramok), |

| | | | |
|---|---|---|---|
| words (n = 3; trigrams) or more are used. | három szóból (n = 3; trigramok) vagy több szóból álló szekvenciákat használunk. | három szóból (n = 3; trigramok) vagy több szóból álló szekvenciákat használunk. | három (n = 3; trigramok) vagy több szóból álló szekvenciákat használunk. |
| 9 The use of higher order ngrams is often optional in tokenization functions. | 9 A magasabb rendű n-grammok használata gyakran opcionális a tokenizáló funkciókban. | A magasabb rendű n-grammok használata gyakran opcionális a tokenizáló funkciókban. | A több tagból álló n-grammok használata gyakran opcionális a tokenizáló funkciókban. |
| quanteda makes it possible to form n-grams when tokenizing, or to form n-grams from tokens already formed. | quanteda lehetővé teszi, hogy a tokenizálás során n-grammokat képezzünk, vagy hogy a már képzett tokenekből n-grammokat képezzünk. | A quanteda lehetővé teszi, hogy a tokenizálás során n-grammokat képezzünk, vagy hogy a már létrehozott tokenekből n-grammokat képezzünk. | A quanteda lehetővé teszi, hogy a tokenizálás során n-grammokat képezzünk, vagy hogy a már létrehozott tokenekből n-grammokat képezzünk. |
| Other options include the formation of “skip-grams”, or n-grams from words with variable windows of adjacency. | Más lehetőségek közé tartozik a "skip-gramok", vagy a változó szomszédsági ablakú szavakból történő n-gramok képzése. | Egy további lehetőség, ha úgynevezett „skip-grammokat” képezünk, azaz olyan, egymás közelében lévő szavak csoportjait hozzuk létre, melyek nem feltétlenül szomszédosak. | Egy további lehetőség, ha úgynevezett kihagyásos n-grammokat [skip-grammokat] képezünk, azaz olyan, egymás közelében lévő szavak csoportjait hozzuk létre, melyek nem feltétlenül szomszédosak. |
| Such non-adjacent collocations form the basis for counting weighted proximity vectors, used in vector-space network-based models built on deep learning techniques (Mikolov, Chen, Corrado, & Dean, 2013; Selivanov, 2016). | Az ilyen nem szomszédos kollokációk képezik a súlyozott közelségvektorok számolásának alapját, amelyeket a mélytanulási technikákra épülő vektortérhálózat-alapú modellekben használnak (Mikolov, Chen, Corrado, & Dean, 2013; Selivanov, 2016). | Az ilyen nem szomszédos szavakból álló kollokációk képezik a súlyozott közelségvektorok számolásának alapját, amelyeket a mélytanulási technikákra épülő vektortérhálózat-alapú modellekben használnak (Mikolov, Chen, Corrado, & Dean, 2013; Selivanov, 2016). | Az ilyen, nem szomszédos szavakból álló kollokációk képezik a súlyozott közelségvektorok kiszámításának alapját. Az alábbiakban bemutatjuk, hogyan lehet képezni 0 és 1-es vektorok felhasználásával mind a trigrammokat, mind a kihagyásokat tartalmazó három elemű n-gramokat 0 és 1-es skipvektorok felhasználásával képezni. |
| The advantage of this approach is that these n-grams can be used in the same way as unigrams: we can make a DTM with n-grams, | Ennek a megközelítésnek az az előnye, hogy ezek az n-grammok ugyanúgy felhasználhatók, mint az unigramok: készíthetünk egy DTM-et n-grammokkal, és | Ennek a megközelítésnek az az előnye, hogy ezek az n-grammok ugyanúgy felhasználhatók, mint az unigramok: készíthetünk egy n-gramokból álló DTM-et, és | Ennek a megközelítésnek az az előnye, hogy az n-grammok ugyanúgy felhasználhatók, mint az unigramok: készíthetünk egy n-gramokból álló DTM-et, és |

| | | | |
|---|---|---|---|
| and perform all the types of analyses discussed above. | elvégezhetjük az összes fent tárgyalt elemzéstípust. | azokon elvégezhetjük az összes fent tárgyalt elemzéstípust. | azokon elvégezhetjük az összes fent tárgyalt elemzéstípust. |
| Functions for creating a DTM in R from raw text therefore often allow the use of n-grams other than unigrams. | Az R-ben a nyers szövegből DTM létrehozására szolgáló függvények ezért gyakran lehetővé teszik az unigramoktól eltérő n-grammok használatát. | Az R-ben a nyers szövegből DTM létrehozására szolgáló funkciók általában lehetővé teszik az unigramoktól eltérő n-grammok használatát is. | Az R-ben a nyers szövegből a DTM létrehozására szolgáló funkciók általában lehetővé teszik az unigramoktól eltérő n-grammok alkalmazását is. |
| For some analysis this can improve performance. | Egyes elemzéseknél ez javíthatja a teljesítményt. | Egyes elemzéseknél ez javíthatja a teljesítményt. | Egyes elemzéseknél ez javíthatja a teljesítményt. |
| Consider, for instance, the importance of negations and amplifiers in sentiment analysis, such as “not good” and “very bad” (Aue & Gamon, 2005). | Gondoljunk például a negációk és erősítők fontosságára a hangulatelemzésben, mint például a "nem jó" és a "nagyon rossz" (Aue & Gamon, 2005). | Gondoljunk például a tagadás és fokozás fontosságára a hangulatelemzésben, mint például a "nem jó" és a "nagyon rossz" (Aue & Gamon, 2005). | Gondoljunk például a tagadás és fokozás fontosságára a hangulatelemzésben, mint például a "nem jó" és a "nagyon rossz" (Aue & Gamon, 2005) kifejezések. |
| An important disadvantage, however, is that using n-grams is more computationally demanding, since there are many more unique sequences of words than individual words. | Fontos hátránya azonban, hogy az n-grammok használata számításigényesebb, mivel sokkal több egyedi szósorozat van, mint egyedi szó. | Fontos hátránya azonban, hogy az n-grammok használata számításigényesebb, mivel sokkal több egyedi szó-sorozat van, mint szó. | A módszer jelentős hátránya azonban, hogy az n-grammok használata számításigényesebb, mert sokkal több egyedi szó-sorozat van, mint szó. |
| Another approach is to preserve the word positions after tokenization. | Egy másik megközelítés a szóhelyzetek megőrzése a tokenizálás után. | Egy másik megközelítés a szóhelyzetek megőrzése a tokenizálás után. | Egy másik lehetséges megközelítés a szóhelyzetek megőrzése a tokenizálás után. |
| This has three main advantages. | Ennek három fő előnye van. | Ennek három fő előnye van. | Ennek három fő előnye van. |
| First, the order and distance of tokens can be taken into account in analyses, enabling analyses such as the co-occurrence of words within a word window. | Először is, a tokenek sorrendjét és távolságát figyelembe lehet venni az elemzések során, ami olyan elemzéseket tesz lehetővé, mint például a szavak egy szóablakon belüli együttes előfordulása. | Egyrészt az elemzések során figyelembe lehet venni a tokenek sorrendjét és távolságát, és ez olyan elemzéseket tesz lehetővé, mint például a szavak egy szóablakon belüli együttes előfordulása. | Egyrészt az elemzések során figyelembe lehet venni a tokenek sorrendjét és távolságát, és ez olyan vizsgálatokat tesz lehetővé, mint például a szavak egy szóablakon (egymáshoz viszonyított, előzetesen definiált távolságán) belüli együttes előfordulása. |
| Second, the data can be transformed into both a fulltext | Másodszor, az adatok átalakíthatók teljes | Másrészt, az adatok átalakíthatók teljes szövegkorpusszá (a tokenek | Másrészt, az adatok átalakíthatók teljes szövegkorpusszá (a tokenek |

| | | | |
|---|---|---|---|
| corpus (by pasting together the tokens) and a DTM (by dropping the token positions). | szövegkorpusszá (a tokenek összeillesztésével) és DTM-mé (a tokenpozíciók elhagyásával). | összeillesztésével) és DTM-mé (a tokenpozíciók elhagyásával). | összeillesztésével) és DTM-mé (a tokenpozíciók elhagyásával). |
| This also enables the results of some text analysis techniques to be visualized in the text, such as coloring words based on a word scale model (Slapin & Proksch, 2008) or to produce browsers for topic models (Gardner et al. , 2010). | Ez lehetővé teszi egyes szövegelemzési technikák eredményeinek szövegben való megjelenítését is, például a szavak színezését egy szóskála-modell alapján (Slapin & Proksch, 2008) vagy a témamodellek böngészőinek előállítását (Gardner et al. , 2010). | Ez lehetővé teszi egyes szövegelemzési technikák eredményeinek szövegben való megjelenítését is, például a szavak színezését egy szóskála-modell alapján (Slapin & Proksch, 2008) vagy a témamodellek böngészőinek előállítását (Gardner et al., 2010). | Ez lehetővé teszi a szövegelemzési technikák eredményeinek megjelenítését is a vizsgált szövegben, ilyen például a szavak szóskála-modell alapján (Slapin & Proksch, 2008) történő színezése vagy a témamodellek generálásához szükséges keresők létrehozása (Gardner et al., 2010). |
| Another approach is to preserve the word positions after tokenization. | Egy másik megközelítés a szóhelyzetek megőrzése a tokenizálás után. | Egy másik megközelítés a szóhelyzetek megőrzése a tokenizálás után. | Egy másik lehetséges megközelítés a szóhelyzetek megőrzése a tokenizálás után. |
| Third, each token can be annotated with token specific information, such as obtained from advanced NLP techniques. | Harmadszor, minden egyes tokent token-specifikus információkkal lehet annotálni, például fejlett NLP-technikákból nyert információkkal. | Harmadrészt minden egyes tokent olyan token-specifikus információkkal lehet annotálni, például fejlett NLP-technikákból nyert információkkal. | Harmadrészt minden egyes tokent olyan token-specifikus, például fejlett NLP-technikák alkalmazásából nyert megjegyzésekkel láthatunk el, |
| This enables, for instance, the use of dependency parsing to perform an analysis at the level of syntactic clauses (Van Atteveldt et al. , 2017). | Ez lehetővé teszi például a függőségi elemzések használatát a szintaktikai záradékok szintjén történő elemzés elvégzéséhez (Van Atteveldt et al., 2017) | Ez lehetővé teszi például a függőségi elemzések használatát a szintaktikai egységek szintjén történő elemzés elvégzéséhez (Van Atteveldt et al. , 2017). | melyek lehetővé teszik például a kapcsolati elemzések használatát a szintaktikai egységek szintjén történő elemzés elvégzéséhez (Van Atteveldt et al. , 2017). |
| The main disadvantage of preserving positions is that it is very memory inefficient, especially if all tokens are kept and annotations are added. | A pozíciók megőrzésének legfőbb hátránya, hogy nagyon memóriahatékony, különösen akkor, ha az összes token megmarad és annotációkat adunk hozzá. | A pozíciók megőrzésének legfőbb hátránya, hogy nagyon nagy memóriát igényel, különösen akkor, ha az összes token megmarad és annotációkat adunk hozzá. | A tokenpozíciók megőrzésének legfőbb hátránya, hogy ez a módszer nagyon nagy memóriát igényel, különösen akkor, ha az összes token megmarad és azokat megjegyzésekkel látjuk el. |
| A common way to represent tokens with positions maintained is a data frame in which rows represent tokens, ordered by their | A pozíciókat fenntartó tokenek ábrázolásának általános módja egy adatkeret, amelyben a sorok a pozíciójuk szerint rendezett | A pozíciókat megőrző tokenek ábrázolásának általános módja egy adatkeret [dataframe], amelyben a sorok a pozíciójuk | A pozíciókat megőrző tokenek ábrázolásának általános módja egy adatkeret [dataframe], amelyben a sorok a pozíciójuk |

| | | | |
|---|---|---|---|
| position, and columns represent different variables pertaining to the token, such as the literal text, its lemma form and its POS tag. | tokeneket, az oszlopok pedig a tokenhez tartozó különböző változókat, például a szó szerinti szöveget, a lemmaformát és a POS-taget ábrázolják. | szerint rendezett tokeneket, az oszlopok pedig a tokenhez tartozó különböző változókat, például az eredeti szerinti szöveget, a lemmaformát és a POS-címkéket ábrázolják. | szerint rendezett tokeneket, az oszlopok pedig a tokenhez tartozó különböző változókat, például az eredeti szerinti szöveget, a lemmaformát és a POS-címkéket ábrázolják. |
| An example of this type of representation was shown above in the advanced NLP section, in the spacyr token output. | Az ilyen típusú reprezentációra fentebb, a fejlett NLP szakaszban, a spacyr token kimeneténél láthattunk példát. | Az ilyen típusú reprezentációra fentebb, a fejlett NLP módszereket tárgyaló szakaszban, a kihagyásos token spacyr token kimeneténél láthattunk példát. | Az ilyen típusú reprezentációra fentebb, a fejlett NLP módszereket tárgyaló szakaszban, a kihagyásos token spacyr token kimeneténél láthattunk példát. |
| Several R packages provide dedicated classes for tokens in this format. | Számos R-csomag biztosít dedikált osztályokat az ilyen formátumú tokenekhez. | Számos R-csomag biztosít dedikált osztályokat az ilyen formátumú tokenekhez. | Számos R-csomag biztosít külön formátumot az ilyen formátumú tokenekhez. |
| One is the koRpus package (Michalke, 2017), which specializes in various types of text statistics, in particular lexical diversity and readability. | Az egyik a koRpus csomag (Michalke, 2017), amely különböző típusú szövegstatisztikákra, különösen a lexikai sokféleségre és az olvashatóságra specializálódott. | Az egyik ilyen a koRpus csomag (Michalke, 2017), amely különböző típusú szövegstatisztikákra, különösen a lexikai sokféleségre és az olvashatóságra specializálódott. | Az egyik ilyen a koRpus program (Michalke, 2017), amely különböző típusú szövegstatisztikákra, különösen a lexikai sokféleségre és az olvashatóság elemzésére kínál megoldásokat. |
| Another is corpustools (Welbers & Van Atteveldt, 2016), which focuses on managing and querying annotated tokens, and on reconstructing texts to visualize quantitative text analysis results in the original text content for qualitative investigation. | Egy másik a corpustools (Welbers & Van Atteveldt, 2016), amely az annotált tokenek kezelésére és lekérdezésére, valamint a szövegek rekonstruálására összpontosít, hogy a kvantitatív szövegelemzés eredményeit az eredeti szövegtartalomban vizualizálja a kvalitatív vizsgálathoz. | Egy másik a corpustools (Welbers & Van Atteveldt, 2016), amely az annotált tokenek kezelésére és lekérdezésére, valamint a szövegek rekonstruálására összpontosít, annak érdekében, hogy a kvantitatív szövegelemzés eredményeit az eredeti szövegtartalomban vizualizálja a kvalitatív vizsgálathoz. | Egy másik a corpustools (Welbers & Van Atteveldt, 2016) program, amely az annotált tokenek kezelésére és lekérdezésére, valamint a szövegek rekonstruálására összpontosít, annak érdekében, hogy a kvantitatív szövegelemzés eredményeit az eredeti szövegtartalomban vizualizálja a kvalitatív vizsgálathoz. |
| A third option is tidytext (Silge & Robinson, 2016), which does not focus on this format of annotated tokens, but provides a | Egy harmadik lehetőség a tidytext (Silge & Robinson, 2016), amely nem az annotált tokenek ilyen formátumára összpontosít, hanem | Egy harmadik lehetőség a tidytext (Silge & Robinson, 2016), amely nem az annotált tokenek ilyen formátumára | Egy harmadik lehetőség a tidytext (Silge & Robinson, 2016). Ez a program nem az annotált tokenekre összpontosít, hanem |

| | | | |
|---|---|---|---|
| framework for working with tokenized text in data frames. | keretrendszert biztosít a tokenizált szöveggel való munkához adatkeretekben. | összpontosít, hanem lehetőséget biztosít a tokenizált szöveggel való munkához adatkeretekben. | lehetőséget biztosít arra, hogy adatkeretekben végezzünk munkát a tokenizált szöveggel. |
| For a brief demonstration of utilizing word positions, we perform a dictionary search with the corpustools package, that supports searching for words within a given word distance. | A szóhelyzetek felhasználásának rövid bemutatására szótárkeresést végzünk a corpustools csomaggal, amely támogatja a szavak keresését egy adott szaktávolságon belül. | A szóhelyzetek felhasználásának rövid bemutatására szó alapú keresést végzünk a corpustools csomaggal, amely támogatja a szavak keresését az adott szók között értelemezett távolságon belül. | Az egyes szavak mondaton belüli pozíviójára alapozott elemzések eredményeinek rövid bemutatására szó alapú keresést végzünk a corpustools programmal, amely támogatja a szavak keresését az adott szók között értelemezett távolságon belül. |
| The results are then viewed in key word in context (KWIC) listings. | Az eredményeket ezután a kulcsszó kontextusban (KWIC) listákban tekintjük meg. | Az eredményeket a kulcsszó kontextusban (KWIC), listák formájában tekintjük meg. | Az eredményeket kulcsszó kontextusban (KWIC), listák formájában tekintjük meg. |
| In the example, we look for the queries “freedom” and “americ*” within a distance of five words, using the State of the Union speeches from George W. Bush and Barack Obama. | A példában a "freedom" és az "americ*" lekérdezéseket keressük öt szó távolságán belül, a George W. Bush és Barack Obama beszédét az Unió helyzetéről. | Bush és Barack Obama elnököknek az USA helyzetéről tartott beszédei alapján. | A példában a "freedom" és az "americ*" lekérdezések eredményeit keressük öt szó távolságán belül, a George W.Bush és Barack Obama elnökök beszédei alapján. |
| Conclusion | Következtetés | Következtetések | Következtetések |
| R is a powerful platform for computational text analysis, that can be a valuable tool for communication research. | Az R egy hatékony platform a számítógépes szövegelemzéshez, amely értékes eszköz lehet a kommunikációkutatásban. | Az R hatékony keretrendszer a számítógépes szövegelemzéshez, amely értékes eszköz lehet a kommunikációkutatásban. | Az R a számítógépes szövegelemzés hatékony keretrendszer, amely a kommunikációkutatás értékes eszköze lehet. |
| First, its well developed packages provide easy access to cutting edge text analysis techniques. | Először is, jól kidolgozott csomagjai könnyű hozzáférést biztosítanak a legmodernebb szövegelemzési technikákhoz. | Először is, jól kidolgozott programjai könnyű hozzáférést biztosítanak a legmodernebb szövegelemzési technikákhoz. | Ez elsősorban azért van így, mert jól kidolgozott programjai könnyű hozzáférést biztosítanak a legmodernebb szövegelemzési technikákhoz. |
| As shown here, not only are most common text analysis techniques implemented, but in most cases, | Amint az itt látható, nemcsak a leggyakoribb szövegelemzési technikák vannak implementálva, | Amint látható, nemcsak a leggyakoribb szövegelemzési technikák vannak implementálva, | Amint bemutattuk, nemcsak a leggyakoribb szövegelemzési eljárásokat építették a |

| | | | |
|---|---|---|---|
| multiple packages offer users choice when selecting tools to implement them. | hanem a legtöbb esetben több csomag is választási lehetőséget kínál a felhasználóknak az eszközök kiválasztásakor. | hanem a legtöbb esetben az eszközök kiválasztásakor több program közül is választhatnak a felhasználók. | programokba, hanem a legtöbb esetben az eszközök kiválasztásakor több program közül is választhatnak a felhasználók. |
| Many of these packages have been developed by and for scholars, and provide established procedures for data preparation and analysis. | E csomagok közül sokat tudósok fejlesztettek ki tudósok számára, és bevált eljárásokat biztosítanak az adatok előkészítéséhez és elemzéséhez. | Ezek közül sokat tudósok fejlesztettek ki tudósok által és tudósok számára, bevált eljárásokat biztosítva az adatok előkészítéséhez és elemzéséhez. | Ezek közül sokat kutatók fejlesztettek ki kutatókkal kutatóknak, bevált eljárásokat biztosítva az adatok előkészítéséhez és elemzéséhez. |
| Second, R's open source nature and excellent system for handling packages make it a convenient platform for bridging the gap between research and tool development, which is paramount to establishing a strong computational methods paradigm in communication research. | Másodszor, az R nyílt forráskódú jellege és a programok kezelésére szolgáló kiváló rendszere kényelmes platformmá teszi a kutatás és az eszközfejlesztés közötti szakadék áthidalására, ami kiemelkedő fontosságú egy megalapozott kvantitatív paradigma kialakításához a kommunikációkutatásban. | Másodszor, az R nyílt forráskódú jellege és a programok kezelésére szolgáló kiváló rendszere kényelmes felületté teszi a kutatás és az eszközfejlesztés közötti szakadék áthidalására, ami kiemelkedő fontosságú egy szilárd kvantitatív módszertan kialakításához a kommunikációkutatásban. | További előny, hogy az R-t nyílt forráskódú jellege és a programok kezelésére szolgáló kiváló rendszere kényelmes felületté teszi a kutatás és az eszközfejlesztés közötti szakadék áthidalására, ami kiemelkedő fontosságú egy szilárd számí kialakításához a kommunikációkutatásban. |
| New algorithms do not have to be confined to abstract and complex explanations in journal articles aimed at methodology experts, or made available through arcane code that many interested parties would not know what to do with. | Az új algoritmusokat nem kell absztrakt és bonyolult magyarázatokba zárni a módszertani szakértőknek szóló folyóiratcikkekben, vagy olyan titkos kódokon keresztül elérhetővé tenni, amelyekkel sok érdeklődő nem tudna mit kezdeni. | Az új algoritmusokat nem kell absztrakt és bonyolult magyarázatokba zárni a módszertani szakértőknek szóló folyóiratcikkekben, vagy olyan titkos kódokon keresztül elérhetővé tenni, amelyekkel sok érdeklődő nem tudna mit kezdeni. | Az új algoritmusokat nem kell absztrakt és bonyolult magyarázatokba zárni a módszertani szakértőknek szóló folyóiratcikkekben, vagy olyan titkos kódokon keresztül elérhetővé tenni, amelyekkel sok érdeklődő nem tudna mit kezdeni. |
| As an R package, algorithms can be made readily available in a standardized and familiar format. | R-csomagként az algoritmusok könnyen hozzáférhetővé tehetők egy szabványosított és ismerős formátumban. | R-csomagként az algoritmusok könnyen hozzáférhetővé tehetők egy szabványosított és megszokott formátumban. | R-csomagként az algoritmusok könnyen hozzáférhetővé tehetők egy szabványosított és megszokott formátumban. |
| For new users, however, choosing from the wide range of text analysis packages in R can also be daunting. | Az új felhasználók számára azonban az R-ben található szövegelemző csomagok széles | Az új felhasználók számára azonban az R-ben található szövegelemző programok széles | A kezdő felhasználók számára már maga az R-ben található szövegelemző programok széles |

| | | | |
|---|---|---|---|
| | választéka közül való választás is ijesztő lehet. | választékából történő választás már maga is ijesztő lehet. | választékából történő választás is ijesztő lehet. |
| With various alternatives for most techniques, it can be difficult to determine which packages are worth investing the effort to learn. | Mivel a legtöbb technikára különböző alternatívák állnak rendelkezésre, nehéz lehet meghatározni, hogy mely csomagok megtanulására érdemes energiát fordítani. | Mivel a legtöbb eljárás megvalósításához különböző alternatívák állnak rendelkezésre, nehéz meghatározni, hogy mely programok megtanulására érdemes energiát fordítani. | A legtöbb eljárás megvalósításához különböző alternatívák állnak rendelkezésre, ezért nehéz meghatározni, hogy mely programok megtanulására érdemes energiát fordítani. |
| The primary goal of this teacher's corner, therefore, has been to provide a starting point for scholars looking for ways to incorporate computational text analysis in their research. | Ennek a tanári saroknak az elsődleges célja ezért az volt, hogy kiindulópontot nyújtson azoknak a tudósoknak, akik a számítógépes szövegelemzés kutatásaikba való beépítésének módját keresik. | Ennek a módszertani segédletnek az volt az elsődleges célja, hogy bevezetést nyújtson azoknak a tudósoknak, akik a számítógépes szövegelemzési kutatásokba szeretnének kezdeni. | Ennek a módszertani segédletnek az volt az elsődleges célja, hogy bevezetést nyújtson azoknak a kutatóknak, akik a számítógépes szövegelemzési feladatokba szeretnének kezdeni. |
| Our selection of packages is based on our experience as both users and developers of text analysis packages in R, and should cover the most common use cases. | A csomagok kiválasztása a szövegelemző R-csomagok felhasználóiként és fejlesztőiként szerzett tapasztalatainkon alapul, és a leggyakoribb felhasználási eseteket hivatott lefedni. | A programok kiválasztása a szövegelemző R-csomagok felhasználóiként és fejlesztőiként szerzett tapasztalatainkon alapul, és a leggyakoribb felhasználási eseteket hivatott lefedni. | A bemutatott programok kiválasztása a szövegelemző R-csomagok felhasználóiként és fejlesztőiként szerzett tapasztalatainkon alapul, és a leggyakoribb felhasználási eseteket hivatott lefedni. |
| In particular, we advise users to become familiar with at least one established and well-maintained package that handles data preparation and management, such as quanteda, tidytext or tm. | Különösen azt tanácsoljuk a felhasználóknak, hogy ismerkedjenek meg legalább egy olyan bevált és jól karbantartott csomaggal, amely az adatok előkészítésével és kezelésével foglalkozik, mint például a quanteda, a tidytext vagy a tm. | Különösen azt tanácsoljuk a felhasználóknak, hogy ismerkedjenek meg legalább egy olyan bevált és jól karbantartott csomaggal, amely az adatok előkészítésével és kezelésével foglalkozik, mint például a quanteda, a tidytext vagy a tm. | Különösen azt tanácsoljuk a felhasználóknak, hogy ismerkedjenek meg legalább egy olyan bevált és jól karbantartott programmal, amely az adatok előkészítésével és kezelésével foglalkozik, mint például a quanteda, a tidytext vagy a tm. |
| From here, it is often a small step to convert data to formats that are compatible with most of the available text analysis packages. | Innen már gyakran csak egy kis lépés az adatokat olyan formátumba konvertálni, amely kompatibilis a legtöbb elérhető szövegelemző csomaggal. | Innen már gyakran csak egy kis lépés az adatokat olyan formátumba konvertálni, amely kompatibilis a legtöbb elérhető szövegelemző csomaggal. | Innen már gyakran csak egy kis lépés az adatokat olyan formátumba történő átalakítása, amely a legtöbb elérhető szövegelemző programmal kompatibilis. |

| | | | |
|---|---|---|---|
| It should be emphasized that the selection of packages presented in this teacher's corner is not exhaustive, and does not represent which packages are the most suitable for the associated functionalities. | Hangsúlyozni kell, hogy az ebben a tanári sarokban bemutatott csomagok kiválasztása nem teljes körű, és nem jelenti azt, hogy mely csomagok a legmegfelelőbbek a kapcsolódó funkciókhoz. | Hangsúlyozni kell, hogy az ebben a módszertani segédletben bemutatott programok kiválasztása nem teljes körű, és nem jelenti azt, hogy mely programok a legmegfelelőbbek a kapcsolódó funkciókhoz. | Hangsúlyozni kell, hogy az ebben a módszertani segédletben bemutatott programok nem fedik le az elérhető szoftverek teljes körét, és nem értelmezehetők úgy, hogy ezek a legmegfelelőbbek a keresett funkciókhoz. |
| Often, the best package for the job depends largely on specific features and problem specific priorities such as speed, memory efficiency and accuracy. | Gyakran az adott feladatra a legjobb csomag kiválasztása nagymértékben függ a konkrét funkcióktól és a problémára jellemző prioritásoktól, például a sebességtől, a memóriahatékonyságtól és a pontosságtól. | Gyakran az adott feladatra a legjobb program választása nagymértékben függ a konkrét funkcióktól és a problémára jellemző prioritásoktól, például a sebességtől, a memóriahatékonyságtól és a pontosságtól. | Az adott feladatra legjobb program megválasztása nagymértékben függ a konkrét funkcióktól és a problémára jellemző prioritásoktól, például a sebességtől, a memóriahatékonyságtól és a pontosságtól. |
| Furthermore, when it comes to establishing a productive workflow, the importance of personal preference and experience should not be underestimated. | Továbbá, amikor egy produktív munkafolyamat kialakításáról van szó, nem szabad alábecsülni a személyes preferenciák és a tapasztalat jelentőségét. | Továbbá, amikor egy produktív munkafolyamat kialakításáról van szó, nem szabad alábecsülni a személyes preferenciák és a tapasztalat jelentőségét. | Amikor egy produktív munkafolyamat kialakításáról van szó, nem szabad alábecsülni a személyes preferenciák és a tapasztalat jelentőségét. |
| A good example is the workflow of the tidytext package (Silge & Robinson, 2016), which could be preferred by people that are familiar with the tidyverse philosophy (Wickham et al.,2014). | Jó példa erre a tidytext csomag munkafolyamata (Silge & Robinson, 2016), amelyet a tidyverse filozófiáját ismerő személyek előnyben részesíthetnek (Wickham et al. , 2014). | Jó példa erre a tidytext program munkafolyamata (Silge & Robinson, 2016), amelyet a tidyverse filozófiáját ismerő személyek előnyben részesíthetnek (Wickham et al., 2014) | Jó példa erre a tidytext program munkafolyamata (Silge & Robinson, 2016), amelyet a tidyverse filozófiáját ismerő személyek előnyben részesíthetnek (Wickham et al., 2014) |
| Accordingly, the packages recommended in this teacher's corner provide a good starting point, and for many users could be all they need, but there are many other great package out there. | Ennek megfelelően az ebben a tanári sarokban ajánlott csomagok jó kiindulópontot jelentenek, és sok felhasználó számára ez lehet minden, amire szüksége van, de sok más nagyszerű csomag is létezik. | Ennek megfelelően az ebben a módszertani segédletben ajánlott programok jó kiindulópontot jelentenek, és sok felhasználó számára ez lehet minden, amire szüksége van, de sok más nagyszerű csomag is létezik. | Ennek megfelelően az ebben a módszertani segédletben ajánlott programok jó kiindulópontot jelentenek, és sok felhasználó számára ez lehet minden, amire szüksége van, de sok más nagyszerű program is létezik. |

| | | | |
|---|---|---|---|
| For a more complete list of packages, a good starting point is the CRAN Task View for Natural Language Processing (Wild, 2017). | A csomagok teljesebb listájához jó kiindulópont a CRAN Task View for Natural Language Processing (Wild, 2017). | A csomagok teljesebb listájához jó kiindulópont a CRAN Task View for Natural Language Processing (Wild, 2017). | A programok teljesebb listájához jó kiindulópont a CRAN Task View for Natural Language Processing (Wild, 2017). |
| Marshall McLuhan famously stated that “we shape our tools, and thereafter our tools shape us”, and in science the same can be said for how tools shape our findings. | Marshall McLuhan híres mondása szerint "mi alakítjuk az eszközeinket, és azután az eszközeink alakítanak minket", és a tudományban ugyanez mondható el arról, hogy az eszközök hogyan alakítják a felfedezéseinket. | Marshall McLuhan híres mondása szerint "mi alakítjuk az eszközeinket, és azután az eszközeink alakítanak minket", és a tudományban ugyanez mondható el arról, hogy az eszközök hogyan alakítják a felfedezéseinket. | Marshall McLuhan híres mondása szerint "mi alakítjuk az eszközeinket, és azután az eszközeink alakítanak minket", és a tudományban ugyanez mondható el arról, hogy az eszközök hogyan alakítják a felfedezéseinket. |
| We thus argue that the establishment of a strong computational methods paradigm in communication research goes hand-in-hand with embracing open-source tool development as an inherent part of scientific practice. | Ezért azt állítjuk, hogy a kommunikációkutatásban egy erős számítási módszertani paradigma kialakítása kéz a kézben jár a nyílt forráskódú eszközfejlesztésnek a tudományos gyakorlat szerves részeként való elfogadásával. | Ezért azt állítjuk, hogy a kommunikációkutatásban egy megalapozott kvantitatív paradigma kialakítása együtt jár azzal, hogy a nyílt forráskódú kutatási eszközök fejlesztése szervesen integrálódik a kutatási gyakorlatba. | Ezért azt állítjuk, hogy a kommunikációkutatásban egy megalapozott kvantitatív paradigma kialakítása együtt jár azzal, hogy a nyílt forráskódú kutatási eszközök fejlesztését a kutatási gyakorlat szerves részének tekintjük. |
| As such, we conclude with a call for researchers to cite R packages similar to how one would cite other scientific work. | Ezért azzal a felhívással zárjuk, hogy a kutatóknak az R-csomagokat hasonlóan kell idézniük, mint ahogyan más tudományos munkákat idéznénk. | Ezért tanulmányunkkat azzal a felhívással zárjuk, hogy a kutatóknak az R-csomagokat hasonlóan kell idézniük, mint ahogyan más tudományos munkákat idéznének. | Ezért tanulmányunkkat azzal a felhívással zárjuk, hogy a kutatóknak az R-csomagokat hasonlóan kell idézniük, mint ahogyan más tudományos munkákat idéznének. |
| This gives due credit to developers, and thereby provides a just incentive for developers to publish and maintain new code, including proper testing and documentation to facilitate the correct use of code by others. | Ez kellő elismerést ad a fejlesztőknek, és ezáltal igazságos ösztönzést nyújt a fejlesztők számára az új kódok közzétételére és karbantartására, beleértve a megfelelő tesztelést és dokumentációt, hogy megkönnyítsék a kód helyes használatát mások számára. | Ez kellő elismerést ad a fejlesztőknek, és ezáltal méltányos ösztönzést ad a fejlesztők számára az új kódok közzétételére és karbantartására, beleértve a megfelelő tesztelést és dokumentációt, hogy megkönnyítsék a kódok helyes használatának mások számára. | Ez kellő elismerést ad a fejlesztőknek, és ezáltal méltányos ösztönzést ad a fejlesztők számára az új kódok közzétételére és karbantartására, beleértve a megfelelő tesztelést és dokumentációt, hogy megkönnyítsék a kódok helyes használatának mások számára. |

| | | | |
|---|---|---|---|
| Citing packages is also paramount for the transparency of research, which is especially important when using new computational techniques, where results might vary depending on implementation choices and where the absence of bugs is often not guaranteed. | A csomagok hivatkozása a kutatás átláthatósága szempontjából is kiemelkedő fontosságú, ami különösen fontos az új számítási technikák alkalmazásakor, ahol az eredmények a megvalósítási választások függvényében változhatnak, és ahol a hibamentesség gyakran nem garantált. | A csomagok hivatkozása a kutatás átláthatósága szempontjából is kiemelkedő fontosságú, ami különösen fontos az új számítási technikák alkalmazásakor, ahol az eredmények a megvalósítási választások függvényében változhatnak, és ahol a hibamentesség gyakran nem garantált. | A csomagok hivatkozása a kutatás átláthatósága szempontjából is kiemelkedő fontosságú. |
| Just as our theories are shaped through collaboration, transparency and peer feedback, so should we shape our tools. | Ahogyan elméleteinket az együttműködés, az átláthatóság és a kollégák visszajelzései alakítják, úgy kell alakítanunk eszközeinket is. | Ahogyan elméleteinket az együttműködés, az átláthatóság és a kollégák visszajelzései alakítják, úgy kell alakítanunk eszközeinket is. | Ez különösen fontos az új számítási technikák alkalmazásakor, ahol az eredmények a megvalósítási választások függvényében változhatnak, és ahol a hibamentesség gyakran nem garantált. |

# Szövegelemzés R-ben

Kasper Welbers[1] , Wouter van Atteveldt[2] & Kenneth Benoit[3]

Institute for Media Studies, University of Leuven[1] , Department of Communcation Science, VU University Amsterdam[2] , Department of Methodology, London School of Economics and Political Science[3]

**Absztrakt**

A számítógépes szövegelemzés izgalmas kutatási területté vált, melynek számos alkalmazása van a kommunikáció kutatásában. Felhasználása bonyolult lehet, mert különböző technikák ismeretét teszi szükségessé, és a legtöbb módszer alkalmazásához szükséges szoftverek nem érhetőek el közvetlenül az általánosan használt statisztikai szoftver programokból. Ebben a módszertani segédletben ezeket az akadályokat tárgyaljuk és átfogó áttekintést adunk egy számítógépes szövegelemzési projekt általános kérdéseiről és műveleteiről, bemutatva, hogy minden egyes lépés hogyan hajtható végre az R statisztikai szoftver felhasználásával. Népszerű, nyílt forrású platformként az R kiterjedt felhasználói közösséggel rendelkezik, mely a szövegelemző programok széles spektrumát fejleszti és tartja fenn. Bemutatjuk, hogy ezen programok hogyan könnyítik meg a bonyolultabb szövegelemzések elvégzését.

A számítógépes szövegelemzés növekvő jelentőségével párhuzamosan a kommunikáció kutatásban számos kutató szembesül azzal a kihívással (Boumans & Trilling, 2016; Grimmer & Stewart, 2013), hogy meg kell tanulnia alkalmazni az ilyen típusú elemzés elvégzésére alkalmas, fejlett szoftvereket. Jelenleg a számítási módszerek és a feltörekvő „adattudomány" egyik legnépszerűbb környezete az R statisztikai szoftver (R Core Team, 2017). A programozásban kevésbé járatos kutatók számára azonban kihívást jelenthet az R használata, és a szövegelemzés megvalósítása különösen ijesztőnek tűnhet. Ebben a módszertani segédletben bemutatjuk, hogy a szövegelemzés végrehajtása nem olyan nehéz, mint amennyire néhányan tartanak tőle. Lépésről lépésre mutatjuk be a leggyakrabban alkalmazott eljárások használatát, azzal a céllal, hogy segítsük a kutatókat a számítógépes szövegelemzés általános megismerésében és az R-ben végzendő fejlett szövegelemzési munkák megvalósításában.

Az R szabadon felhasználható, nyílt forráskódú, platformfüggetlen programozási környezet. Más programnyelvekkel ellentétben az R-t kifejezetten statisztikai elemzésre tervezték, ezért kiválóan alkalmas adattudományi alkalmazásokra. Bár az R programozásának tanulási görbéje – különösen azok számára, akik nem rendelkeznek programozási tapasztalatokkal – meredek lehet, az R-ben történő szövegelemzés megvalósítását szolgáló, jelenleg rendelkezésre álló eszközök megkönnyítik a mindössze néhány egyszerű parancs alkalmazásával végrehajtott hatékony, korszerű szövegelemzést. Az R robbanásszerű bővülésének (Fox & Leanage, 2016; TIOBE, 2017) egyik kulcsa a nagyszámú kiegészítő szoftverkönyvtár, az úgynevezett programok, melyeket az R kiterjedt felhasználói közössége készít és tart fenn. Minden egyes program tovább bővíti az R-nyelv és az alapprogramok felhasználási lehetőségeit. Az új funkciók és adatok hozzáadásához dokumentációra és példákra is szükség van, melyek gyakran címkék formájában mutatják be a program alkalmazását. A legismertebb program-könyvtár az átfogó R-archívum [Comprehensive R

Archive Network (CRAN)], mely jelenleg több mint 10 000 közzétett, az eljárások összhangja és a platformfüggetlenség szempontjából előzetesen és széleskörűen ellenőrzött programot tartalmaz. Így az R kiterjedt, egymással komptatibilis programokból áll, melyeket a tudósok, a gyakorló szakemberek és az olyan projektek, mint például az Rstudio és az rOpenSci folyamatosan karbantartanak és frissítenek. Fontos tudnunk, hogy ezeket a programokat könnyen és biztonságosan, egy parancs alkalmazásával lehet installálni az R-környezetből. Ebből adódóan az R szilárd, a fejlesztők és az új elemzési eszközök felhasználói közötti találkozási felületet képez, ezért ez a programozási környezet rendkívül alkalmas a tudományos együttműködésre.

A szövegelemzés különösen jól illeszkedik az R-be. A legutóbbi felmérésünk alapján összesen közel ötven programot magába foglaló, kifejezetten a szövegfeldolgozást és a szövegelemzést szolgáló hatalmas gyűjteménye van, az alacsony szintű karakterlánc-kezelési műveletektől (Gagolewski, 2017) az olyan magas szintű szöveg-modellezési eljárásokig, mint például a látens Dirichlet allokációs modellek (Blei, Ng, & Jordan, 2003; Roberts et al., 2014). A fejlesztők között fokozódó erőfeszítések tapasztalhatók az együttműködésre és az erőfeszítések összehangolására, ilyen például az rOpenSci szakosított együttműködés. Az R-ben végrehajtott szövegelemzés egyik fő előnye, hogy gyakran lehetséges és viszonylag könnyű a különböző programok közötti váltás, vagy azok együttes felhasználása. A R szövegelemzésfejlesztő közösségének legutóbbi erőfeszítési arra irányulnak, hogy elősegítsék az egyes programok közötti interoperabilitást a rugalmasság és a felhasználó szabadságának maximalizálása érdekében. Így, ha elsajátítjuk az R-ben végzett szövegelemzés alapjait, a fejlettebb szövegelemzési lehetőségek egész sora nyílik meg előttünk.

## A Módszertani Segédlet felépítése

Ez a módszertani segédlet az R-ben végzett leggyakoribb szövegelemzési lépéseket tartalmazza, az adat-előkészítéstől az elemzésig, és minden egyes lépéshez könnyen megismételhető példa-kódokat nyújt. A példa-kódok digitálisan is elérhetők a folyamatosan frissített online függelékben is. Elsősorban a szózsák típusú szövegelemzési megközelítésre összpontosítunk. Ez azt jelenti, hogy csak a szövegenként előforduló szavak gyakoriságát használjuk fel és nem vesszük figyelembe azok helyzetét. Ez erőteljesen egyszerűsíti ugyan a szöveg tartalmát, de a kutatások és számos gyakorlati alkalmazás azt mutatja, hogy a szavak gyakorisága önmagában elégséges információt tartalmaz sokféle elemzés elvégzéséhez (Grimmer & Stewart, 2013).

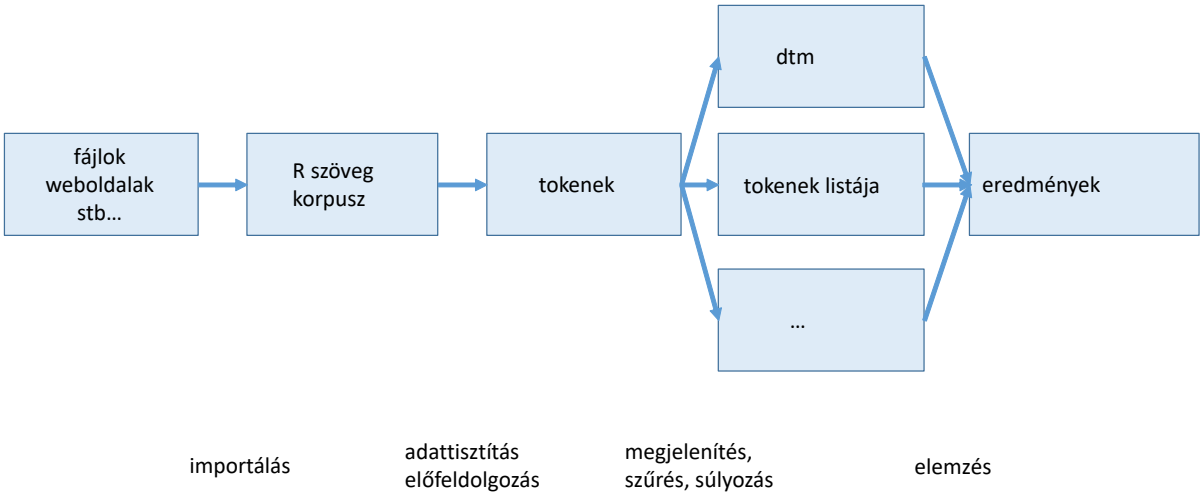
Az 1. táblázatban áttekintés adunk a bemutatott szövegelemzési műveletekről, három csoportra osztva azokat. Az *adatelőkészítésről* szóló fejezetben az elemzésre történő szövegelőkészítés öt lépését tárgyaljuk. Az első lépés a szöveg *importálása*, mely a szöveg különböző file-formátumokból (pl. txt, csv, pdf) a *nyers szöveg-testbe* (korpuszba) történő beolvasásához szükséges feladatokat foglalja magába az R-be. Ezek a karakterlánc-műveletek és az előzetes feldolgozás azokat a nyers szövegkezelési és előfeldolgozási technikákat jelentik, melyek célja a nyers szöveg átalakítása és feldolgozása *tokenekké* (szótövekké) (azaz szövegelemekké, például szavakká és szótövekké). Ezt követően a példányokat a *dokumentum-kifejezés mátrix* (document-term matrix, DTM) létrehozására használjuk fel. Ez általánosan használt formája a szózsák-típusú szövegtörzsek megjelenítésének, melyeket számos R szövegelemző program használ. Egyéb, nem szózsák alapú formátumokat, például a példánylistákat ugyancsak bemutatunk a"*bonyolultabb kérdések*" fejezetben. Végül gyakori az egyes kifejezések *szűrése és súlyozása* is a DTM-ben (document-term matrix).

1. táblázat

A szövegelemzéshez alkalmazott módszerek és az azokhoz a jelen Módszertani Segédletben felhasznált R programok áttekintése

| **Művelet** | **R programok** | |
|---|---|---|
| | példa | alternatívák |
| **Adatelőkészítés,** szöveg-importálás | readtext | jsonlite, XML, antiword, readxl, pdftools |
| karakterműveletek | stringi | stringr |
| előzetes feldolgozás | quanteda | stringi, tokenizers, snowballC, tm, etc. |
| document-kifejezés mátrix (DTM) | quanteda | tm, tidytext, Matrix |
| szűrés és súlyozás | quanteda | tm, tidytext, Matrix |
| **Elemző** műveletek | quanteda | tm, tidytext, koRpus, corpustools |
| Felügyelt gépi tanulás | quanteda | RTextTools, kerasR, austin |
| nem felügyelt gépi tanulás | topicmodels | quanteda, stm, austin, text2vec |
| szövegstatisztika | quanteda | koRpus, corpustools, textreuse |
| **Bonyolultabb feladatok** | | |
| fejlett természetes nyelv-feldolgozási módszerek | spacyr | coreNLP, cleanNLP, koRpus |
| a szavak helyzete és a szintaktikai elemzés | corpustools | quanteda, tidytext, koRpus |

1. ábra A szövegelemzés műveleteinek sorrendje az adatfeldolgozástól az analízisig




Ezeket a lépéseket általában a bemutatott sorrendben hajtjuk végre (a fogalmi illusztrációhoz lásd az 1. ábrát). Amint látni fogjuk, vannak R-programok, melyek megfelelő funkciókkal rendelkeznek ahhoz, hogy többféle adatelőkészítő lépés egyetlen programsorral megvalósítható legyen. Ennek ellenére mi először minden egyes lépést külön-külön tárgyalunk és mutatunk be, annak érdekében, hogy alapszinten megértsük az egyes lépések célját, a választási lehetőségeket és felhívjuk a figyelmet a buktatókra.

Az *elemzés* fejezetben a kommunikációkutatásban népszerű négy szövegelemzési módszert tárgyalunk (Boumans & Trilling, 2016), melyeket a DTM-input felhasználásával hajthatunk végre. Ahelyett, hogy ezeket egymással versengő megközelítéseknek tekintenénk, látnunk kell, hogy az egyes módszereknek különböző előnyeik és hátrányaik vannak, és ezért a kutatás szempontjából legjobb módszer kiválasztása nagymértékben függ a kutatási kérdéstől, és néha a különböző módszerek kiegészíthetik egymást (Grimmer & Stewart, 2013). Ennek megfelelően azt javasoljuk, hogy mindegyik módszerrel érdemes megismerkedni. Annak érdekében, hogy bemutassuk, milyen általános elvre épülnek az egye módszertípusok, a tipikus elemzési példák kódját is megadjuk. Vegyük figyelembe azt is, hogy a különböző elemzések eltérő adatelőkészítést igényelhetnek. Ebből adódóan minden egyes elemzési típus esetében az adatelőkészítéssel kapcsolatos megfontolásokat is tárgyaljuk.

Végül, a *bonyoltabb témakörök* fejezetben alternatív adat-előkészítési és elemzési módszereket tárgyalunk, melyek külső szoftver-modulok alkalmazását teszik szükségessé vagy túlmennek a szózsák-feltételezésen, mert felhasználják a szavak helyzetét és szintaktikai kapcsolatait is. Ennek a fejezetnek az a célja, hogy betekintést nyújtson az R-rel elérhető alternatív, olykor bonyolultabban használható megoldásokra. Minden egyes kategórián belül néhány műveletcsoportot különítünk el, és minden egyes műveletre bemutatjuk, hogy az hogyan valósítható meg R-ben. Annak érdekében, hogy tömör és egyszerűen megismételhető példákat mutassunk be, olyan programokat választottunk, melyek egyszerűen

használhatók és széles körben alkalmazhatók. Tudnunk kell azonban, hogy számos alternatív program is elérhető, melyek azonos vagy hasonló műveletek végrehajtására képesek. Az R nyílt forráskód jellegéből adódóan a különböző kutatók gyakran egymásól eltérő tudományterületekről dolgoznak hasonló problémákon, és így az egyes funkciók között átfedések alakulhatnak ki az egyes programok esetén. Ez egyúttal a felhasználói igénynek és ízlésnek megfelelő alternatívákat kínál. A kutatási projekttől és az egyéni preferenciáktól függően lehetséges, hogy más programok jobban megfelelnek a különböző olvasóknak. Jóllehet az R szövegelemző programok teljeskörű áttekintése és összehasonlítása meghaladja vizsgálataink körét – különösen azért, mert folyamatosan fejlesztik a meglévő és az új programokat – mégis megkíséreltük legalább az említés szintjén érinteni az egyes szövegelemzési műveletek során alkalmazható alternatív programokat. Általában ezen programok gyakran használják ugyanazt az standard adatformátumot, és így könnyen helyettesíthetjük vagy kombinálhatjuk azokat a jelen módszertani segédletben bemutatott programokkal.

## A dokumentum-term mátrix

A dokumentum-term mátrix (DTM) a szöveges állományok (lényegében szövegek gyűjteményei) megjeltítésének egyik leggyakoribb módja a szövegzsák formátumban. A DTM mátrixban a sorok jelentik a dokumentumokat, az oszolopok a teminusokat, a cellák pedig azok gyakoriságát az egyes dokumentumokban. Ennek a megjelenítésnek az az előnye, hogy lehetővé teszi az adatok statisztikai elemzését, azaz a szövegről a számokra történő áttérést. A ritka mátrixok speciális mátrixformátumainak használatával a DTM -ben tárolt szöveges adatok nagyon hatékony memóriakezelést és optimalizált műveletvégzést tesznek lehetővé.

Az tm és a quanteda az R két legelterjedtebb szövegelemző programja, amelyekkel speciális DTM fájlok hozthatók létre A kettő közül a jóöreg tm program a leggyakrabban alkalmazott, közel 10 éves felhasználói tapasztalattal (Meyer, Hornik, & Feinerer, 2008). Az általa létrehozott DTM fájlokat (*DocumentTermMatrix* és *TermDocumentMatrix*) számos más R program is használja az elemző funkcióik bemeneteként. A quanteda program újabb fejlesztésű, amelyet egy ERC-ösztöndíjjal támogatott csapat azért hozott létre, hogy a legkorszerűbb, nagy teljesítményű szövegelemzést valósítsa meg. Az programmal létrehozott, a ritka mátrixok kezelését lehetővé tevő DTM fájl, amelyet dfm vagy dokumentum-tulajdonságmátrix néven ismerünk, háttértárként a nagy teljesítményű Matrix programra (Bates & Maechler, 2015) épül, de tartalmaz olyan funkciókat is, amelyekkel szinte minden más, R-programban használt ritka dokumentum-term mátrix (beleértve a tm formátumokat is) is konvertálható. A quanteda `dfm` fájlját teljesítménye és rugalmassága miatt ennek használatát ajánljuk a tm megfelelő funkciójával szemben.

Egy további figyelemre méltó alternatíva a tidytext program (Silge & Robinson, 2016). Ez egy olyan szövegelemző program, amely része a *Tidyversenek*, azaz egy közös elven és formátumon alapuló R-programok gyűjteményének. A Tidyverse filozófiájának központi eleme, hogy minden adat táblázatba rendeződik, ahol (1) "minden változó egy oszlopot", (2) "minden megfigyelési egység egy sort" és (3) "minden adatgyűjtés-sorozat egy táblázatot" alkot (Wickham et al. , 2014, 4). Mint ilyen, a tidytext nem szigorúan dokumentum-terminus mátrixot használ, hanem ugyanezeket az adatokat egy olyan hosszú formátumban tárolja, ahol a DTM minden (nem nulla) értéke egy sor, és melynek oszlopai a dokumentum, a terminus és a gyakoriság (megjegyzendő, hogy ez lényegében a ritka mátrixok triplett formátuma, ahol az oszlopokba renezett információk adják meg a sort, az oszlopot és az értéket). Ez a formátum kevésbé hatékony a memóriakezelés szempontjából, és a mátrixalgebra is nehezebben alkalmazható, de előnye, hogy több változót (pl. egy érzelmi értékelést) adhat hozzá a táblázathoz, és lehetővé teszi a teljes Tidyverse-eszköztár használatát. Így a tidy adatkezelési elveket kedvelő felhasználók számára a tidytext jó

alternatíva lehet a quanteda vagy a tm mellett, bár ezek a csomagok a kívánt konkrét műveletektől függően együtt is nagyon hatékonyan használhatók.

A módszertani segédlet többi példájával összhangban a DTM-ek létrehozását a quanteda program segítségével mutatjuk be. Ennek `dfm` funkciója a korábban bemutatott előfeldolgozási technikákat integrálva egyetlen lépésben nyújt megoldást a DTM létrehozására nyers szövegből.

```
text <- c(d1 = "An example of preprocessing techniques",
d2 = "An additional example",
d3 = "A third example")
dtm <- dfm(text,                          # szövegbevitel
tolower = TRUE, stem = TRUE,              # átaállítás kisbetűre és szótő-keresésre, ha TRUE-eltávolítja az
remove = stopwords("english"))            # angol stopwordsöket
dtm
```

```
Document-feature matrix of: 3 documents, 5 features (53.3% sparse).
3 x 5 sparse Matrix of class "dfmSparse"
features
docs exampl preprocess techniqu addit third
 d1      1          1         1      0     0
 d2      1          0         0      1     0
 d3      1          0         0      0     1
```

A DTM egy quanteda korpuszobjektumból is létrehozható, amely a szöveget és a kapcsolódó metaadatokat tárolja, ideértve a dokumentumszintű változókat is. Amikor egy korpuszt tokenizálunk vagy DTM-mé alakítunk át, ezek a dokumentumszintű változókat az adott fájlba mentjük el. Ez később nagyon hasznos lehet, amikor a DTM-ben lévő dokumentumokat kovariátorként (moderator vagy mediator változók) kell használnunk a felügyelt gépi tanulás során. A tárolt dokumentumváltozók lehetővé teszik a quanteda fájlok csoportok szerinti aggregálását is, ami rendkívül hasznos akkor, ha a szövegek kis egységekben - például tweetekben - vannak tárolva, de a DTM-ben olyan változók szerinti csoportosításban kell aggregálnunk őket, mint például a felhasználók, a dátumok vagy ezek kombinációi.

A quanteda kompatibilis a pkgreadtext programmal, a korpusz létrehozása a tárolt szövegekből csak egyetlen további lépést igényel. A következő példában a fentiek szerint importált readtext adatokból hozunk létre egy DTM-et.

```
fulltext <- corpus(rt)                                    # quanteda korpuszt hozunk létre
```

```
dtm <- dfm(fulltext, tolower = TRUE, stem = TRUE,          # előfeldolgozással dtm-et alkotunk
 with preprocessing remove_punct = TRUE,remove = stopwords("english"))
dtm
```

```
Document-feature matrix of: 5 documents, 1,405 features (67.9% sparse).
```

**Szűrés és súlyozás**

Nem minden kifejezés egyformán informatív a szövegelemzés szempontjából. Ezt úgy lehet kezelni, hogy ezeket a kifejezéseket eltávolítjuk a DTM-ből. A nagyon gyakori kifejezések eltávolítására már tárgyaltuk a stopword-listák használatát, de általában vannak még más gyakori szavak is, arányuk viszont az az egyes korpuszok között eltérő lesz. A nagyon ritka kifejezések eltávolítása is hasznos lehet az olyan feladatokhoz, mint például az automatikus osztályzos (Yang & Pedersen, 1997) vagy a témamodellezés (Griffiths & Steyvers, 2004). Ez különösen hasznos a hatékonyság javítása szempontjából, mert nagymértékben csökkentheti a korpuszban lévő egyedi szavak és kifejezések számát, sőt a pontosságot is javíthatja. Egyszerű, de hatékony módszer a dokumentumfrekvenciák (azon dokumentumok száma, amelyekben egy kifejezés előfordul) alapján történő szűrés, a dokumentumok minimális és maximális számának (vagy arányának) küszöbértékét használva (Griffiths & Steyvers, 2004; Yang & Pedersen, 1997).

A kevésbé informatív kifejezések eltávolítása helyett egy alternatív megközelítés lehet az, ha változó súlyokat rendelünk hozzájuk. Számos szövegelemzési eljárás jobban teljesít, ha a kifejezések súlyozása egy becsült információérték figyelembevételével történik, ahelyett, hogy közvetlenül az előfordulási gyakoriságukat használnánk. Kellően nagy korpusz esetén a kifejezések eloszlásával kapcsolatos információkat felhasználhatjuk a korpuszban az információs érték becsléséhez. A terminusfrekvencia-inverz dokumentumfrekvencia (*tf-idf*) egy olyan elterjedt súlyozási séma, amely a korpusz sok dokumentumában előforduló kifejezéseket kisebb súlyszám-értékkel veszi figyelembe.

A dokumentumfrekvencia küszöbérték és súlyozás használata egyszerűen elvégezhető a DTM-en. A quanteda tartalmazza a `docfreq`, `tf` és `tfidf` funkciókat a dokumentumfrekvencia, a terminusfrekvencia és a *tf-idf* kiszámításához. Mindegyik funkció számos lehetőséget kínál a SMART súlyozási séma megvalósításához (Manning et al., 2008). A quanteda ezekhez átfogó, magas szintű lehetőségként a `dfm_weight` funkciót is biztosítja. Az alábbi példában a "senat[e]" szónak nagyobb súlya van, mint a kevésbé informatív "among" kifejezésnek, amelyek egyszer-egyszer fordulnak elő az első dokumentumban.

```
doc_freq <- docfreq(dtm)          # dokumentumgyakoriság terminusonkétn (oszlopban)
dtm <- dtm[, doc_freq >= 2]       # terminusválasztás doc_freq >= 2 dtm <- dfm_weight(dtm, "tfidf")
                                  # a jellemzők súlyozása tf-idf használattal(dtm)
```

```
Document-feature matrix of: 5 documents, 6 features (40% sparse).
```

```
5 x 6 sparse Matrix of class "dfmSparse" features
```

| docs | fellow-citizen | senat | hous | repres | among | life |
|---|---|---|---|---|---|---|
| text1 | 0.2218487 | 0.39794 | 0.79588 | 0.4436975 | 0.09691001 | 0.09691001 |
| text2 | 0 | 0 | 0 | 0 | 0 | 0 |
| text3 | 0.6655462 | 0.39794 | 1.19382 | 0.6655462 | 0.38764005 | 0.19382003 |
| text4 | 0.4436975 | 0 | 0 | 0.2218487 | 0.09691001 | 0.09691001 |
| text5 | 0 | 0 | 0 | 0 | 0.67837009 | 0.19382003 |

**Elemzés**

A szövegelemzési megközelítések áttekintéséhez a Boumans és Trilling (2016) által javasolt osztályozásra építünk, amelyben három módszert különböztet meg: számláló és szótári módszerek, felügyelt gépi tanulás és felügyelet nélküli gépi tanulás. Ezek a módszereket ebben a sorrendben a leginkább deduktívtól a leginkább induktívig terjedő koordinátatengely mentén helyezkednek el. A deduktív ebben az esetben egy előre (*a priori*) meghatározott kódolási séma használatára utal. Más szóval a kutatók előre tudják, hogy mit keresnek, és csak az elemzés automatizálására törekszenek. A kapcsolat a deduktív érvelés fogalmával az, hogy a kutató feltételezi, hogy bizonyos szabályok vagy premisszák igazak (pl. a pozitív hangulatot jelző szavak listája), és így alkalmazhatók a szövegekre vonatkozó következtetések levonására. Ezeket alacsonyabb szintű funkciók sorozatán keresztül is fel lehet építeni, de sok felhasználó számára kényelmes, ha egy szövegből vagy korpuszból közvetlenül jut el a DTM-hez. A DTM egy quanteda korpuszobjektumból is létrehozható, amely a szöveget és a kapcsolódó metaadatokat tárolja, ideértve a dokumentumszintű változókat is. Amikor egy korpuszt tokenizálunk vagy DTM-mé alakítunk át, ezek a dokumentumszintű változókat az adott fájlba mentjük el. Ez később nagyon hasznos lehet, amikor a DTM-ben lévő dokumentumokat kovariátorként (moderator vagy mediator változók) kell használnunk a felügyelt gépi tanulás során. A tárolt dokumentumváltozók lehetővé teszik a quanteda fájlok csoportok szerinti aggregálását is, ami rendkívül hasznos akkor, ha a szövegek kis egységekben - például tweetekben - vannak tárolva, de a DTM-ben olyan változók szerinti csoportosításban kell aggregálnunk őket, mint például a felhasználók, a dátumok vagy ezek kombinációi. Az induktív ezzel szemben itt azt jelenti, hogy egy a priori kódolási séma használata helyett a számítógépes algoritmus maga határozza meg az értelmes kódokat a szövegekből. Ez például úgy történhet, hogy mintázatokat keres a szavak együttes előfordulásában, és olyan látens tényezőket (pl. témák, keretek, szerzők) talál, amelyek ezeket a mintákat - legalábbis matematikailag - megmagyarázzák. Az induktív érvelés szempontjából azt mondhatjuk, hogy az algoritmus konkrét megfigyelések alapján széleskörű általánosításokat hoz létre.

E három kategória mellett egy statisztikai kategóriát is figyelembe veszünk, amely magában foglalja a szöveg vagy korpusz számokkal történő leírására szolgáló összes technikát. A felügyelet nélküli tanuláshoz hasonlóan ezek a technikák is induktívak abban az értelemben, hogy sem a priori kódolási sémát, sem gépi tanulást nem alkalmaznak.

Az egyes elemzési típusokat bemutató minta-parancsokhoz az amerikai elnökök- a quanteda programban szereplő- beiktatási beszédeit (N = 58) használjuk.

```
dtm <- dfm(data_corpus_inaugural, stem = TRUE, remove = stopwords("english"),
remove_punct = TRUE)

dtm
```

```
## Document-feature matrix of: 58 documents, 5,405 features (89.2% sparse).
```

**Számlálás és szótárak**

A tágabb értelemben vett szótári megközelítés a mintázatok felhasználására utal - az egyszerű kulcsszavaktól az összetett Bool-algebrai-lekérdezésekig és a kötött kifejezésekig, állandósult szókapcsolatokig –, annak érdekében, hogy megszámoljuk, milyen gyakran fordulnak elő bizonyos fogalmak a szövegekben. Ez deduktív megközelítés, mert a szótár a priori meghatározza, hogy milyen kódokat és hogyan kell mérni, és ezt az adatok nem befolyásolják. A szótárak használata számítási szempontból egyszerű, de hatékony módszer. Olyan témák tanulmányozására használták, mint például a politikai szereplőkre irányuló médiafigyelem (Schuck, Xezonakis, Elenbaas, Banducci és De Vreese, 2011; Vliegenthart, Boomgaarden és Van Spanje, 2012), vagy a vállalati információk kontextusba foglalása (Schultz, Kleinnijenhuis, Oegema, Utz és Van Atteveldt, 2012). A szótárak szintén népszerű eszköznek számítanak az érzelmek (De Smedt & Daelemans, 2012; Mostafa, 2013; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011), valamint a szubjektív nyelvhasználat más dimenzióinak mérésére (Tausczik & Pennebaker, 2010). Az ilyen típusú elemzések és a szintaktikai egységek azonosítására szolgáló fejlett NLP-technikákból származó információk kombinálásával részletesebb elemzések elvégzése is lehetővé válik, például az egyes szereplőknek tulajdonított érzelemkifejezések (Van Atteveldt, 2008) vagy az egyik szereplőtől a másik felé irányuló cselekvések és érzelmek vizsgálata (Van Atteveldt, Sheafer, Shenhav, & Fogel-Dror, 2017).

A következő példa azt mutatja be, hogyan alkalmazható egy szótár egy quanteda DTM-re. Az első lépés egy szótár fájl létrehozása (itt myDict), a dictionary funkció használatával. Az egyszerűség kedvéért példánk egy nagyon egyszerű szótárat használ, de lehetőség van nagy, előre elkészített szótárak importálására is, ideértve más szövegelemző szótár-formátumú fájlokat is, mint például a LIWC, Wordstat és a Lexicoder. A szótárak YAML fájlokból is írhatók és importálhatók, és tartalmazhatnak fix egyezésekből, szabályos kifejezésekből vagy az egyszerűbb "glob" szerkezetekből álló mintázatokat, (csak * és ? karaktereket használva joker karakterekként), amelyek számos szótárformátumban elterjedtek. A dfm_lookup funkcióval a szótárobjektumot egy DTM-re lehet alkalmazni egy új DTM létrehozásához, amelyben a szótárkódokat az oszlopok tartalmazzák.A felügyelt gépi tanulás módszere minen olyan kutatási technikát magába foglal, amikor egy algoritmus mintákat tanul egy további információkkal kiegészített gyakorló adathalmazból.

```
myDict <- dictionary(list(terror = c("terror*"), economy = c("job*", "business*", "econom*")))

dict_dtm <- dfm_lookup(dtm, myDict, nomatch = "_unmatched") tail(dict_dtm)
```

Document-feature matrix of: 6 documents, 3 features (16.7% sparse).

6 x 3 sparse Matrix of class "dfmSparse" features

| docs | terror | economy | _unmatched |
|---|---|---|---|
| 1997-Clinton | 2 | 3 | 1125 |
| 2001-Bush | 0 | 2 | 782 |
| 2005-Bush | 0 | 1 | 1040 |
| 2009-Obama | 1 | 7 | 1165 |
| 2013-Obama | 0 | 6 | 1030 |
| 2017-Trump | 1 | 5 | 709 |

**Felügyelt gépi tanulás**

A felügyelt gépi tanulás módszere minen olyan kutatási technikát magába foglal, amikor egy algoritmus mintákat tanul egy további információkkal kiegészített gyakorló adathalmazból. Az intuitív módszer úgy írható le, hogy ezek az algoritmusok megtanulják, hogyan kell kódolni a szövegeket, ha arra elegendő példát adunk nekik. Egyszerű példa erre a hangulatelemzés, amely egy ember által pozitívnak, semlegesnek vagy negatívnak kódolt szövegkészletet használ. Ennek alapján az algoritmus megtanulhatja, hogy mely jellemzők (szavak vagy szóösszetételek) fordulnak elő nagyobb valószínűséggel pozitív vagy negatív szövegekben. Egy ismeretlen szöveg segítségével (amelyre az algoritmust nem tanították be) a szöveg hangulatát ezeknek a jellemzőknek a jelenléte alapján meg lehet becsülni. A deduktív fázisban a kutatók biztosítják a tanuláshoz szükséges jó példákat tartalmazó adatokat, kiegészítve azokat a megfelelő kategóriákkal. A kutatók azonban nem adnak explicit szabályokat arra vonatkozóan, hogy hogyan keressük ezeket a kategóriákat. Az induktív szakaszban a felügyelt gépi tanulási algoritmus a tanítására felhasznált adatokból tanulja meg ezeket a szabályokat. Egy klasszikus szillogizmust parafrazálva: ha a képzési adatok olyan emberek listája, akik vagy halandóak, vagy halhatatlanok, akkor az algoritmus megtanulja, hogy minden ember nagy valószínűséggel halandó, és így azt is megállapítaná, hogy Szókratész is halandó.

Ennek bemutatására egy modellt tanítottunk be annak előrejelzésére, hogy egy [amerikai elnöki] beiktatási beszédet a II. Világháború előtt vagy után mondtak-e el, aminek sikerességére azért számítunk, mert – különösen a világháború után – változtak a kiemelt témák idővel. A felügyelt gépi tanuláshoz többféle program áll rendelkezésre, pl.az RTextTools (Jurka, Collingwood, Boydstun, Grossman, & Van Atteveldt, 2014) és a kerasR (Arnold, 2017b). Ehhez a példához a quanteda egyik oszályozó algoritmusát használunk. Mielőtt elkezdenénk, beállítunk egy egyéni kiinduló értéket az R véletlenszám-generátorának, hogy a kód véletlenszerű részeinek eredményei mindig ugyanazok legyenek. Az adatok előkészítéséhez hozzáadjuk a DTM-hez az is_prewar dokumentum (meta) változót, amely jelzi, hogy mely dokumentumok származnak 1945 előttről. Ez az a változó, amelyet a modellünk megpróbál előre jelezni. Ezután a DTM-et tanulási (`train_dtm`) és teszt (`test_dtm`) adatokra osztjuk, a tanuláshoz 40 dokumentumból álló véletlenszerű mintát, a teszteléshez pedig a maradék 18 dokumentumot használjuk. A tanulási adatokat egy multinomiális naív Bayes módszeren alapuló klasszifikációs algoritmus (Manning et al., 2008,13. fejezet) tanítására használjuk, amelyet az nb_modelhez rendelünk. A modell klasszifikációs képességének vizsgálatára, azaz annak eldöntésére, hogy egy beiktatásib beszéd a háború előtt vagy után hangzott-e el, az alogrimtussal elkésszíttetjük a tesztadatokra alapozott előrejelzést, majd készítünk egy táblázatot, amelyben a sorok az előrejelzést, az oszlopok pedig az `is_prewar` változó tényleges értékét mutatják.

```
set.seed(2)
# create a document variable indicating pre or post war
docvars(dtm, "is_prewar") <- docvars(dtm, "Year") < 1945
# sample 40 documents for the training set and use remaining (18) for testing
train_dtm <- dfm_sample(dtm, size = 40)
test_dtm <- dtm[setdiff(docnames(dtm), docnames(train_dtm)), ]
# fit a Naive Bayes multinomial model and use it to predict the test data
 nb_model <- textmodel_NB(train_dtm, y = docvars(train_dtm, "is_prewar"))
 pred_nb <- predict(nb_model, newdata = test_dtm)
# compare prediction (rows) and actual is_prewar value (columns) in a table
table(prediction = pred_nb$nb.predicted, is_prewar = docvars(test_dtm, "is_prewar"))
```

```
is_prewar
prediction FALSE TRUE
         FALSE 8          0
         TRUE   0          10
```

Az eredmények azt mutatják, hogy az előrejelzés tökéletes. Mind a nyolc alkalom közül, amikor a program, a HAMIS állítást jelezte előre (azaz a beszéd a világháború előtt hangzott el), és mind a tíz alkalommal, amikor az IGAZ állítást jelezte előre, ez volt a valóság.

**A nem felügyelt tanulás**

A felügyelet nélküli gépi tanulásnál nincsenek sem előre meghatározott kódolási szabályok, sem további információkat tartalmazó adatok a betanításhoz. Ehelyett egy algoritmus a szövegben található mintázatok azonosításával hoz létre modellt. A kutató egyetlen befolyásolási lehetősége bizonyos paraméterek megadása, például a képzendő csoportok száma.Az ilyen algoritmusok használatának széleskörűen alkalmazott példája a témamodellezés, amikor dokumentumokat automatikus osztályozunk egy, még nem ismert témastruktúra alapján (Blei et al., 2003; Roberts et al. , 2014)-, és a "Wordfish" parametrikus faktor modellt (Proksch & Slapin, 2009) használjuk a dokumentumok egyetlen dimenzió, például a bal-jobb ideológia alapján történő skálázására.

Grimmer és Stewart (2013) szerint a felügyelt és a felügyelet nélküli gépi tanulás nem konkurens módszerek, hanem különböző feladatokat látnak el, és nagyon is jól kiegészíthetik egymást. A felügyelt módszerek alkalmazása akkor a legcélszerűbb, ha a dokumentumokat előre meghatározott kategóriákba kell sorolni, mert nem valószínű, hogy egy felügyelet nélküli módszer olyan kategorizálást fog eredményezni, amely megfelel ezeknek a kategóriáknak és lehetővé teszi a kategorizálás kutatói értelmezését. A nem felügyelt módszerek némileg kiszámíthatatlan természetének előnye, hogy olyan kategóriák is létrejöhetnek, amelyekre a kutatók nem gondoltak. (Amikor az eredmények nem egyértelműek, ez kihívást jelenthet az utólagos értelmezés során.) A felügyelet nélküli tanulás lényegének bemutatására az alábbi példa szolgál, mely azt demonstrálja, hogyan lehet illeszteni egy témamodellt az R-ben a topicmodels program (Grun & Hornik, 2011) segítségével.

Ahhoz, hogy pontosabban öszpontosíthassunk a beiktatási beszédekben felmerülő témákra, és hogy növeljük a modellezendő szövegek számát, először bekezdésenként felosztjuk a szövegeket, és létrehozunk egy új DTM-et. Ebből a DTM-ből eltávolítjuk az öt, vagy annál kisebb gyakoriságú kifejezéseket, hogy csökkentsük a szavak számát (ez a jelenlegi példa szempontjából kevésbé fontos), és a quanteda convert funkciójával a DTM-et a topicmodels által használt formátumba konvertáljuk. Ezután egy vanílla LDA témamodellt (Blei et al. , 2003) tanítunk be öt témával. Az LDA nem determinisztikus, ezért az eredmények reprodukálhatósága érdekében rögzített kiindulási szimulációs értékeket használunk.

```
install.packages("topicmodels") library(topicmodels) texts =
corpus_reshape(data_corpus_inaugural, to = "paragraphs")

par_dtm <- dfm(texts, stem = TRUE,      # create a document-term matrix

remove_punct = TRUE, remove = stopwords("english"))

par_dtm <- dfm_trim(par_dtm, min_count = 5)   # remove rare terms

par_dtm <- convert(par_dtm, to = "topicmodels") # convert to topicmodels format

set.seed(1)

lda_model <- topicmodels::LDA(par_dtm, method = "Gibbs", k = 5) terms(lda_model, 5)
```

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|---|
| [1,] | "govern" | "nation" | "great" | "us" | "shall" |
| [2,] | "state" | "can" | "war" | "world" | "citizen" |
| [3,] | "power" | "must" | "secur" | "new" | "peopl" |
| [4,] | "constitut" | "peopl" | "countri" | "american" | "duti" |
| [5,] | "law" | "everi" | "unit" | "america" | "countri" |

Az eredmények az egyes témákhoz kapcsolódó terminusokat mutatják be. Jóllehet, ez csak egy egyszerű példa, mégis jól látszik, hogy alulról felfelé történő közelítéssel öt témakör rajzolódikki az egyes témákhoz sorolt terminusok szemantikai koherenciája alapján. Látható például, hogy az egyik témakör a "govern[ance]", "power", "state", "constitut[ion]", és a "law" szavak köré szerveződik.

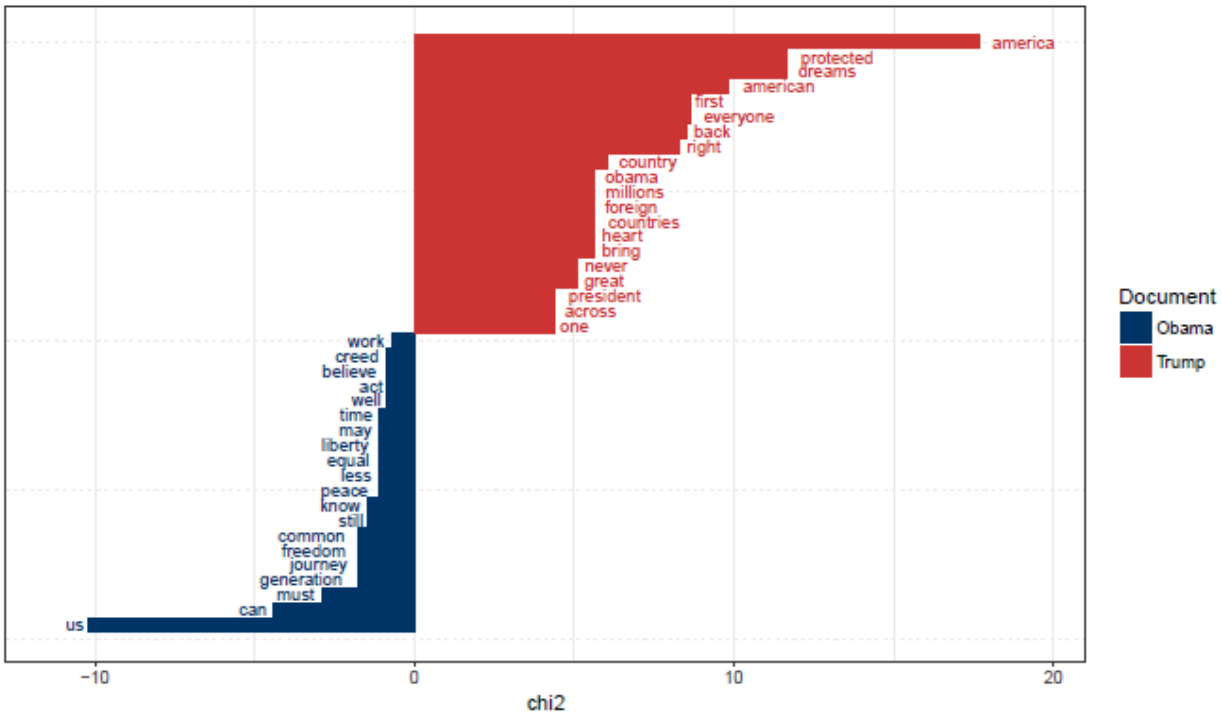
**Statisztikák**

A szövegkorpuszok leírására, feltárására és elemzésére különböző statisztikák használhatók. Gyakran alkalmazott eljárás például a korpuszban található szavak információértékének rangsorolása, majd a leginformatívabb szavak szófelhő formájában történő megjelenítése annak érdekében, hogy gyors képet kapjunk a korpusz tartalmáról. A szövegstatisztikákat (pl. átlagos szó- és mondathossz, szó- és szótagszám) szintén gyakran használják az olyan fogalmak operacionalizálására, mint például az olvashatóság (Flesch, 1948) vagy a lexikai sokszínűség (McCarthy & Jarvis, 2010). Az ilyen mérőszámok széles skálája elérhető az R-ben a koRpus csomaggal (Michalke, 2017). A terminus- és dokumentumhasonlóságok kiszámításának (amelyek gyakran egy DTM vagy transzponált DTM belső szorzatán alapulnak) számos felhasználási lehetősége van, például a szavak vagy fogalmak közötti szemantikai kapcsolatok elemzése és a tartalmi homogenitás mérése. Mindkét technikát támogatja a quanteda, és a corpustools (Welbers & Van Atteveldt, 2016) továbbá a szövegek átfedésének vizsgálatára szolgáló speciális csomagok, például a

textreuse (Mullen, 2016a). Különösen hasznos, ha két korpuszban, vagy ugyanazon korpusz két részében hasonlítjuk össze az egyes terminusok gyakoriságát. Ilyenkor például megvizsgáljuk, hogy mely szavak fordulnak elő nagyobb gyakorisággal egy adott témájú dokumentumban. Ez amellett, hogy lehetőséget nyújt annak gyors feltárására, hogy az adott témát hogyan tárgyalják a korpuszban, ötleteket adhat hatékonyabb lekérdezések kidolgozásához.

A következő példában bemutatjuk, hogyan lehet ezt az eljárást a quanteda-ban végrehajtani (Eredményeinket a 2. ábra szemlélteti).

2. *ábra* A legfőbb szavak előfordulásának relatív gyakorisága Trum és Obama elnökök beszédeiben




```
# create DTM that contains Trump and Obama speeches
corpus_pres = corpus_subset(data_corpus_inaugural,
                            President %in% c("Obama", "Trump"))
dtm_pres = dfm(corpus_pres, groups = "President",
               remove = stopwords("english"), remove_punct = TRUE)

# compare target (in this case Trump) to rest of DTM (in this case only Obama).
keyness = textstat_keyness(dtm_pres, target = "Trump")
textplot_keyness(keyness)
output in Figure 2.
```

Itt a χ2-val jeölt asszociációs mérőszám azt mutatja, hogy az "america", "american" és "first" szavakat Trump sokkal gyakrabban használta, mint Obama, míg az "us", "can", "freedom", "peace" és "liberty" szavakat Obama használta sokkal gyakrabban, mint Trump.

## ÖSSZETETT MÓDSZEREK

A fentiekben tárgyalt adatelőkészítési és szóhalmazelemzési módszerek képezik a kommunikációkutatásban jelenleg alkalmazott a szövegelemzési módszerek többségének az alapját.

Bizonyos típusú elemzésekhez azonban szükség lehet olyan technikákra, amelyek külső szoftvermodulokra támaszkodnak, számításigényesebbek vagy használatuk bonyolultab. Ebben a fejezetben röviden bemutatunk néhány ilyen fejlett módszert, amelyeket érdemes figyelembe venni.

Az adatelőkészítéssel foglalkozó fejezetben tárgyalt előfeldolgozási módszereken kívül vannak olyan hatékony előfeldolgozási eljárások is, amelyek a fejlettebb, természetes nyelvi feldolgozásra (NLP) támaszkodnak. Jelenleg ezek a módszerek nem állnak rendelkezésre az R-ben, hanem külső szoftvermodulokra támaszkodnak, amelyeket az R-en kívül kell telepíteni. A fejlett NLP-technikák közül sok számításigényesebb is, így a végrehajtásuk több időt vesz igénybe. További nehézség, hogy ezek az eljárások nyelvspecifikusak, és gyakran csak az angol és néhány más nagy nyelvhez kapcsolódnak.

Néhány R csomag biztosít interfészeket külső NLP modulokhoz, így ha ezeket a modulokat egyszer már telepítettük, könnyen használhatjuk őket R-ből. A coreNLP csomag (Arnold & Tilton, 2016) kapcsolatot biztosít a Stanford CoreNLP java könyvtárhoz (Manning et al., 2014), amely egy teljes körű NLP elemző az angol nyelvhez, amely (bár korlátozásokkal) támogatja az arab, kínai, francia, német és spanyol nyelveket is. A spacyr szoftver (Benoit & Matsuo, 2017) interfészt biztosít a python spaCy moduljához, amely a CoreNLP-hez hasonló, de gyorsabb, és támogatja az angol nyelvet, valamint (szintén bizonyos korlátozásokkal) a németet és a franciát is. Egy harmadik szoftver, a cleanNLP (Arnold, 2017a) kényelmesen csomagolja mind a CoreNLP-t, mind a spaCy-t, és tartalmaz egy minimális back-endet is, amely nem támaszkodik külső függőségekre. Így svájci bicskaként használható, az alkalomhoz legjobban illő megközelítést választva, amelyhez a back-end rendelkezésre áll, de egységesített kimenettel és módszerekkel. Néhány R program biztosít interfészeket külső NLP modulokhoz, így ha ezeket a modulokat egyszer már telepítettük, könnyen használhatjuk őket R-ből.

A coreNLP program (Arnold & Tilton, 2016) kapcsolatot biztosít a Stanford CoreNLP java könyvtárhoz (Manning et al., 2014), amely egy teljes körű NLP elemző szoftver az angol nyelvhez, és amelyik (korlátozásokkal ugyan) támogatja az arab, kínai, francia, német és spanyol nyelveket is. A spacyr szoftver (Benoit & Matsuo, 2017) interfészt biztosít a python spaCy moduljához, amely a CoreNLP-hez hasonló, de gyorsabb, és támogatja az angol nyelvet, valamint (szintén bizonyos korlátozásokkal) a németet és a franciát is. Egy harmadik szoftver, a cleanNLP (Arnold, 2017a) kényelmesen integrálja mind a CoreNLP-t, mind a spaCy-t, és tartalmaz egy minimális háttérszámításokat végző molult is, amely nem támaszkodik külső szoftverekre. Így egyfajta "svájci bicskaként" használható, a feladathoz legjobban illő módszert választva, amelyhez a háttérszámításokat végző modul egységesített kimenettel és módszerekkel áll rendelkezésre.

## Rövidítések jegyzéke

| Rövidítés | angol jelentés | magyar jelentés |
|---|---|---|
| CRAN | Comprehensive R Archive Network | Átfogó R archívum könyvtár |
| CSV | comma separated values | vesszővel tagolt értékek |
| DTM | Document-term matrix | Dokumentum-kifejezés mátrix/dokumentum-terminus mátrix |
| GNU | GNU's Not UNIX | „a GNU nem Unix” |
| HTML | Hyper Text Markup Language | Hipertext-alapú jelölőnyelv |
| JSON | JavaScript Object Notation | Java szkript alapú objektum jelölés |
| NLP | Natural Language Professing | Természetes nyelvfeldoglozás |
| NLP | Natural Language Processing | Természetes nyelvfeldoglozás |
| PDF | Portable document file | Hordozható dokumentumfájl |
| TXT | Text file | szövegfájl |
| URL | Uniform resource locator | egységes erőforrás-helymeghatározó |
| UTF | Unicode Transformation Format | Unikód átalakítási forma |
| XML | Extensible Markup Language | kiterjeszthető jelölőnyelv |
| YAML | Yet another markup language | Még egy jelölőnyelv |

## FELHASZNÁLT IRODALOM


Boda, Zsolt, and Sebők, Miklós. 2018. A magyar közpolitikai napirend: Elméleti alapok, empirikus eredmények. Magyar Tudományos Akadémia Társadalomtudományi Kutatóközpont

Demeczky, Jenő. 2008. "Terminológia a szoftveriparban." *Magyar Terminológia* no. 1 (2):189-204.

Fóris, Ágota. 2020. *Lexikológiai és lexikográfiai ismeretek*: Károli Gáspár Református Egyetem, L'Harmattan Kiadó.

Hargittay, E. 2016. *Pázmány Péter írói módszere és az életmű genezise: a Kalauz és a vitairatok újraírása, Akadémiai Doktori Értekezés*.

Jockers, Matthew L, and Rosamond Thalken. 2020. *Text analysis with R*: Springer.

Mészáros, Evelin, and Miklós Sebők. 2018. A szövegbányászati módszerek alkalmazásának lehetőségei a joggyakorlat-elemzésben. Paper read at Forum Sententiarum Curiae.

MTA. 2015. *A magyar helyesírás szabályai*. Budapest: Akadémiai Kiadó.

Magyar Országgyűlés. 2023. "2023. évi XXIII. törvény a kiberbiztonsági tanúsításról és a kiberbiztonsági felügyeletről."

Porkoláb, Ádám, and Fekete, Tamás. 2021. *Magyar-angol és angol-magyar nyelvészeti szakszótár*. Pécs: Pécsi Tudományegyetem.

Ring, Orsolya, and Kiss, László. 2020. "Agrárpolitikai kihívások és jogszabályalkotás a korai Kádár-korban: Történeti források szövegelemzése és szövegbányászati vizsgálata." *Digitális Bölcsészet* (3):37-T: 61.

Strelicz, Andrea. 2022. "A kvantitatív szövegelemzés lehetőségei a menedzsmenttudományban= Possibilities of quantitative text analysis in management science." *STATISZTIKAI SZEMLE* no. 100 (6):551-583.

Szita, Szilvia. 2020. "Korpuszépítés és korpuszhasználat alacsonyabb nyelvtudási szinteken." *Hungarológiai Évkönyv* no. 21 (1-2):173-179.