Sebők–Ring–Máté

Szövegbányászat és mesterséges intelligencia R-ben

Sebők Miklós
Ring Orsolya
Máté Ákos

Szövegbányászat és mesterséges intelligencia R-ben

TYPOTEX

A kézirat előkészítését és a kötet megjelenését támogatta:

Társadalomtudományi Kutatóközpont – MTA Kiváló Kutatóhely

Eötvös Loránd Kutatási Hálózat

MTA Könyvkiadási Alap

MTA Bolyai János Kutatási Ösztöndíj

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal (NKFIH FK 123907, NKFIH FK 129018)

A kötet alapjául szolgáló kutatást, amelyet a Társadalomtudományi Kutatóközpont valósított meg, az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.

ISBN 978 963 493 139 3

Kedves Olvasó!
Köszönjük, hogy kínálatunkból választott olvasnivalót!
Újabb kiadványainkról és akcióinkról a *www.typotex.hu*
és a *facebook.com/typotexkiado* oldalakon értesülhet.

Typotex Kiadó
Alapította Votisky Zsuzsa, 1989
A kiadó az 1795-ben alapított Magyar Könyvkiadók
és Könyvterjesztők Egyesülésének tagja.
Felelős kiadó: Németh Kinga
Főszerkesztő: Horváth Balázs
A kötetet gondozta: Erő Zsuzsa
Olvasószerkesztő: Fedinec Csilla
Szakmai lektor: Ballabás Dániel
Borítóterv: Szalay Éva
Készült a Multiszolg Bt. nyomdájában
Felelős vezető: Kajtor Bálint

A borítón a Hungarian Comparative Agendas Project napirend előtti felszólalásokat tartalmazó, 2006–2010-es időszakot lefedő korpuszán elvégzett wordfish elemzés vizualizációja látható.
A vizualizációt készítette: Székely Anna

Tartalomjegyzék

1. fejezet

Bevezetés

1.1. A kötet témái

A szövegek adatként való értelmezése (*text as data*) és kvantitatív elemzése (*quantitative text analysis*), avagy a szövegbányászat (*text mining*) a nemzetközi társadalomtudományi kutatások egyik leggyorsabban fejlődő irányzata. A szövegek és más kvalitatív adatok (filmek, képek) elemzése annyiban különbözik a mennyiségi (kvantitatív) adatokétól, hogy nyers formájukban még nem alkalmasak statisztikai, illetve ökonometriai elemzésre. Ezért van szükség az ezzel összefüggő módszertani problémák speciális tárgyalására.

Jelen kötet bevezeti az érdeklődőket a szövegbányászat és a mesterséges intelligencia társadalomtudományi alkalmazásának ilyen speciális problémáiba, valamint ezek gyakorlati megoldásába. Közvetlen előzménynek tekinthető a témában a Sebők Miklós által szerkesztett *Kvantitatív szövegelemzés és szövegbányászat a politikatudományban* címmel megjelent könyv, amely a magyar tudományos diskurzusban kevésbé bevett alapfogalmakat és eljárásokat mutatta be (Sebők, 2016). A hangsúly az elméleten volt, bár számos fejezet foglalkozott konkrét kódrészletek elemzésével. Míg az előző kötet az egyes kódolási eljárásokat, illetve ezek kutatásmódszertani előnyeit és hátrányait ismertette, ezúttal a társadalomtudományi elemzések során használható kvantitatív szövegelemzés legfontosabb gyakorlati feladatait vesszük sorra.

Könyvünk a magyar tankönyvpiacon elsőként ismerteti lépésről-lépésre a nemzetközi társadalomtudományban használatos kvantitatív szövegelemzési eljárásokat. A módszereink bemutatására szolgáló elemzéseket az R programozási nyelv segítségével végeztük el, mely a nemzetközi társadalomtudományi vizsgálatok során egyik leggyakrabban használt környezet a Python mellett. A kötetben igyekeztünk magyar szakkifejezéseket használni, de mivel a szövegbányászat nyelve az angol, mindig megadtuk azok angol megfelelőjét is. Kivételt képeznek azok az esetek, ahol nincs használatban megfelelő magyar terminológia, ezeknél megtartottuk az angol kifejezéseket, de magyarázattal láttuk el azokat.

Az Olvasó a két kötet együttes használatával olyan ismeretek birtokába jut, melyek révén képes lesz alkalmazni a kvantitatív szövegelemzés és szövegbányászat legalapvetőbb eljárásait saját kutatásaiban. Deduktív vagy induktív felfedező logikája szerint dönthet az adatelemzés módjáról, és a felkínált menüből kiválaszthatja a kutatási tervéhez legjobban illeszkedő megoldásokat. A bemutatott konkrét példák segítségével pedig akár reprodukálhatja is ezen eljárásokat saját kutatásában. Mindezt a kötet fejezeteiben bőséggel tárgyalt *R scriptek* (kódok) részletes leírása is segíti. Ennek alapján a kötet két fő célcsoportja a társadalomtudományi kutatói és felsőoktatási hallgatói-oktatói közösség. Az oktatási alkalmazást segítheti a tárgymutató, valamint több helyen a további olvasásra ajánlott szakirodalom felsorolása. A kötet honlapján (`tankonyv.poltextlab.com`) közvetlenül is elérhetőek a felhasznált adatbázisok és kódok.

Kötetünk négy logikai egységből épül fel. Az első négy fejezet bemutatja azokat a fogalmakat és eljárásokat, amelyek elengedhetetlenek egy szövegbányászati kutatás során, valamint itt kerül sor a szöveges adatforrásokkal való munkafolyamat ismertetésére, a szövegelőkészítés és a korpuszépítés technikáinak bemutatására. A második blokkban az egyszerűbb elemzési módszereket tárgyaljuk, így a leíró statisztikák készítését, a szótáralapú elemzést, valamint érzelemelemzést. A kötet harmadik blokkját a mesterséges intelligencia alapú megközelítéseknek szenteljük, melynek során az olvasó a felügyelt és felügyelet nélküli tanulás fogalmával ismerkedhet meg. A felügyelet nélküli módszerek közül a topikmodellezést, szóbeágyazást és a szövegskálázás wordfish módszerét mutatjuk be, a felügyelt elemzések közül pedig az osztályozással foglalkozunk részletesebben. Végezetül kötetünket egy függelék zárja, melyben a kezdő RStudio felhasználóknak adunk gyakorlati iránymutatást a programfelülettel való megismerkedéshez, használatának elsajátításához.

1.2. Használati utasítás

A könyv célja, hogy keresztmetszeti képet adjon a szövegbányászat R programnyelven használatos eszközeiről. A fejezetekben ezért a magyarázó szövegben maga az R kód is megtalálható, illetve láthatóak a lefuttatott kód eredményei. Az alábbi példában a sötét háttér az R környezetet jelöli, ahol az R kód betűtípusa is eltérő a főszövegétől. A kód eredményét pedig a `#>` kezdetű sorokba szedtük, ezzel szimulálva az R console ablakát.

```
# példa R kód
1 + 1
#> [1] 2
```

Az egyes fejezetekben szereplő kódrészleteket egymás utáni sorrendben bemásolva és lefuttatva a saját R környezetünkben tudjuk reprodukálni a könyvben szereplő technikákat. A Függelékben részletesebben is foglalkozunk

az R és az RStudio beállításaival, használatával. Az ajánlott R minimum verzió a 4.0.0, illetve az ajánlott minimum RStudio verzió az 1.4.0000.[1]

A könyvhöz tartozik egy `HunMineR` nevű R csomag is, amely tartalmazza az egyes fejezetekben használt összes adatbázist, így az adatbeviteli problémákat elkerülve lehet gyakorolni a szövegbányászatot. A könyv megjelenésekor a csomag még nem került be a központi R CRAN csomag repozitóriumába, hanem a poltextLAB GitHub repozitóriumából tölthető le.

A könyvben szereplő ábrák nagy része a `ggplot2` csomaggal készült a `theme_set(theme_light())` opció beállításával a háttérben. Ez azt jelenti, hogy az ábrákat előállító kódok a `theme_light()` sort nem tartalmazzák, de a tényleges ábrán már megjelennek a tematikus elemek.

Csomagnév	Verziószám
dplyr	1.0.5
e1071	1.7.6
factoextra	1.0.7
gapminder	0.3.0
GGally	2.1.1
ggdendro	0.1.22
ggplot2	3.3.3
ggrepel	0.9.1
HunMineR	0.0.0.9000
igraph	1.2.6
kableExtra	1.3.4
knitr	1.33
lubridate	1.7.10
purrr	0.3.4
quanteda	3.0.0
quanteda.textmodels	0.9.4
quanteda.textplots	0.94
quanteda.textstats	0.94
readr	1.4.0
readtext	0.80
readxl	1.3.1
rvest	1.0.0
spacyr	1.2.1
stm	1.3.6
stringr	1.4.0
text2vec	0.6
tibble	3.1.1
tidyr	1.1.3
tidytext	0.3.1
topicmodels	0.2.12

1.1. táblázat. A könyvben használt R csomagok

[1] Az R Windows, macOS és Linux változatai itt érhetőek el: `https://cloud.r-project.org/`
Az RStudio pedig innen érhető el: `https://www.rstudio.com/products/rstudio/download/`

Az egyes fejezetekben használt R csomagok listája és verziószáma az 1.1 táblázatban található. Fontos tudni, hogy a használt R csomagokat folyamatosan fejlesztik, ezért elképzelhető hogy eltérő verziószámú változatok esetén változhat a kódszintaxis. Egyes függvényeket az első megjelenésükkor a `csomag::függvény` szintaxissal használjuk, hogy egyértelmű legyen, hogy melyik csomag része az adott függvény.

1.3. A HunMineR használata

A Windows rendszert használóknak először az `installr` csomagot kell telepíteni, majd annak segítségével letölteni az Rtools nevű programot (az macOS és Linux rendszerek esetében erre a lépésre nincs szükség). A lenti kód futtatásával ezek a lépések automatikusan megtörténnek.

```
# az installr csomag letöltése és installálása
install.packages("installr")

# az Rtools.exe fájl letöltése és installálása
installr::install.Rtools()
```

Ezt követően a `devtools` csomagban található `install_github` paranccsal tudjuk telepíteni a `HunMineR` csomagot, a lenti kód lefuttatásával.

```
# A devtools csomag letöltése és installálása
install.packages("devtools")

# A HunMineR csomag letöltése és installálása
devtools::install_github("poltextlab/HunMineR")
```

Ebben a fázisban a `data` függvénnyel tudjuk megnézni, hogy pontosan milyen adatbázisok szerepelnek a csomagban, illetve ugyanitt megtalálható az egyes adatbázisok részletes leírása. Ha egy adatbázisról szeretnénk többet megtudni, akkor a kiegészítő információkat `?adatbazis_neve` megoldással tudjuk megnézni.[2]

```
# A HunMineR csomag betöltése
library(HunMineR)

# csomagban lévő adatok listázása
data(package = "HunMineR")

# A miniszterelnöki beszédek minta adatbázisának részletei
?data_miniszterelnokok
```

[2] Többek között az adat forrása, a változók részletes leírása, illetve az adatbázis mérete is megtalálható így.

1.4. Köszönetnyilvánítás

Jelen kötet az ELKH Társadalomtudományi Kutatóközpont poltextLAB szövegbányászati kutatócsoportja (`http://poltextlab.com/`) műhelyében készült. A kötet fejezetei Sebők Miklós, Ring Orsolya és Máté Ákos közös munkájának eredményei. Az *Alapfogalmak*, illetve a *Szövegösszehasonlítás* fejezetekben társszerző volt Székely Anna. A kézirat a szerzők többéves oktatási gyakorlatára, a hallgatóktól kapott visszajelzésekre építve készült el. Köszönjük a Bibó Szakkollégiumban (2021), a Rajk Szakkollégiumban (2019–2021), valamint a Széchenyi Szakkollégiumban (2019) tartott féléves, valamint a Corvinus Egyetemen és a Társadalomtudományi Kutatóközpontban tartott rövidebb képzési alkalmak résztvevőinek visszajelzéseit. Köszönjük a projekt gyakornokainak, Czene-Joó Máténak, Kaló Eszternek, Meleg Andrásnak, Lovász Dorottyának, Nagy Orsolyának, valamint a kutatás asszisztenseinek, Balázs Gergőnek, Gelányi Péternek és Lancsár Eszternek a kézirat végleges formába öntése során nyújtott segítséget.

Külön köszönet illeti a Társadalomtudományi Kutatóközpont Comparative Agendas Project (`https://cap.tk.hu/hu`) kutatócsoportjának tagjait, kiemelten Boda Zsoltot, Molnár Csabát és Pokornyi Zsanettet a kötetben használt korpuszok sokéves előkészítéséért. Köszönettel tartozunk az egyes fejezetek alapjául szolgáló elemzések és publikációk társszerzőinek, Barczikay Tamásnak, Berki Tamásnak, Kacsuk Zoltánnak, Kubik Bálintnak, Molnár Csabának és Szabó Martina Katalinnak.

Köszönjük Ballabás Dániel szakmai lektor hasznos megjegyzéseit, Fedinec Csilla nyelvi lektor alapos munkáját, valamint a Typotex Kiadó rugalmasságát és színvonalas közreműködését a könyv kiadásában! Végül, de nem utolsósorban hálásak vagyunk a kötet megvalósulásához támogatást nyújtó szervezeteknek és ösztöndíjaknak: az MTA Könyvkiadási Alapjának, a Társadalomtudományi Kutatóközpont Könyvtámogatási Alapjának, a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatalnak (NKFIH FK 123907, NKFIH FK 129018), az MTA Bolyai János Kutatási Ösztöndíjának.

A kötet alapjául szolgáló kutatást, amelyet a Társadalomtudományi Kutatóközpont valósított meg, az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.