Kvantitatív szövegelemzés

2019 SZISZ - őszi félév

Instruktor: Máté Ákos

Email: aakos.mate@gmail.com

Kurzus leírása

A modern társadalomtudomány/adatbányászat egyik új módszertani innovációja a kvantitatív szövegelemzés, ami az eddig nehezen, drágán (vagy sehogy) elemezhető szövegek feldolgozását teszi lehetővé automatizált eszközökkel. A kurzus célja, hogy megismertesse a kvantitatív szövegelemzés elméleti és gyakorlati oldalát a résztvevőkkel és átfogó képet adjon a területről. Ennek részeként megismerkedünk a főbb kapcsolódó kutatásmódszertani alapelvekkel, az egyes algoritmusokkal és az alkalmazásaikkal.

A hangsúly a gyakorlati oktatáson van, a félév során a megismert módszereket R-ben fogjuk alkalmazni. Miért R? A Python mellett az egyik legelterjedtebb programozási nyelv az adatelemzési feladatokhoz, amit az akadémián és privát szférában egyaránt széleskörben használnak és a szövegelemzési képességei dinamikusan fejlődnek.

A kurzusnak nem előkövetelménye sem az R, sem bármely másik programozási nyelv ismerete (de kétségtelenül előnyt jelent). Mivel a hangsúly az elméletek és módszerek alkalmazásán van, ezért különösebb matematikai/statisztikai előtanulmányok sem szükségesek.

Előkövetelmény - Középszintű szakmai angol nyelvtudás

Követelmények

1. Órai jelenlét és részvétel (15%) A cél hogy a problémákat megbeszélve, a kis csoportlétszámot kihasználva tudjunk haladni. A kurzus szemináriumjellegű, minden felvetődött kérdést, problémát meg lehet beszélni.

2. Négy beadandó feladat (egyenként 5%, összesen 20%)

A feladatok gyakorlás jellegűek, céljuk hogy rendszeres motivációt jelentsenek az R használatához, illetve hogy a megismert módszertani eszköztárat magabiztosan tudják a kurzus végére használni a hallgatók.

3. Projekt terv prezentáció (10%)

A félév végi kutatási projekt terv prezentációja. Tartalmaznia kell a kutatási problémát, az alkalmazott módszertant illetve a felhasználandó adatok bemutatását. Kb 10-15 perces hosszúságú szóbeli előadások.

4. Projekt feladat (55%)

Egy szabadon választott projekt feladat, ami a kurzuson tanult módszertant hasznosítja. Terjedelem: 3000 szó (kb 17 000 karakter). A maximum terjedelme a beadandónak 3500 szó (hivatkozásokat nem számolva). A használt adatok lehetnek saját gyüjtésű illetve már létező adatbázisban lévő szövegek, a kutatási kérdésben, illetve alkalmazott módszertanban megkötés nincs azon túl, hogy a kurzuson tanultakat kell hasznositani. A választott témának nincs megkötése, Jane Austen korai munkássága és Donald Trump twitterje között bármilyen téma belefér. A beadandó részletes formája:

```
- Bevezetés és kutatási kérdés (kb 500 szó)
- Adat és módszertan (kb 800 szó)
- Elemzés (kb 1200 szó)
- Konklúzió (kb 500 szó)
```

Heti tematika

Hét 1 - Kvantitativ Szövegelemzés alapjai.

Bevezető alkalom, alapfogalmak tisztázása, a terület áttekintése

Kötelező olvasmány:

- Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21, no. 3 (2013): 267-297.

Hét 2 - Bevezetés az R-be

Ismerkedés az R-el, adatbevitel és manipuláció, szövegek bevitele. (ha lesz rá idő) Adatvizualizació a ggplot2-vel és dokumentumok készítése az RMarkdown-al.

Kötelező olvasmány:

- Wickham, Hadley, and Garrett Grolemund. *R for data science: import, tidy, transform, visualize, and model data.*, O'Reilly Media, Inc., 2016., **27. fejezet** (online elérhető: https://r4ds.had.co.nz/)
- Healy, Kieran. *Data visualization: a practical introduction.* Princeton University Press, 2018., **1. fejezet** (online elérhető: http://socviz.co/lookatdata.html#lookatdata)

Hét 3 - Szöveg mint adat 1.

A szövegek adatként kezelése, leiró statisztikák készitése, adat preparáció a későbbi elemzéshez.

Kötelező olvasmány:

- Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2), 168-189.
- Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. *CommunicationMethods and Measures*, 11(4), 245-265.

Hét 4 - Szöveg mint adat 2.

A szövegek adatként kezelése, leiró statisztikák készitése, szófrekvencia alapú elemzés

Kötelező olvasmány:

- Krippendorff, Klaus. 2004. Content Analysis: An Introduction to Its Methodology., Chapter 9

Hét 5 - Szótár alapú módszerek

Szentiment elemzés, szótárak konstrukciója és limitációi.

Kötelező olvasmány:

- Laver, M., & Garry, J. (2000). Estimating policy positions from political texts. American Journal of Political Science, 619-634.
- Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. Political Communication, 29(2), 205-231.

Hét 6 - Machine learning applikációk szövegek klasszifikációjára.

A Naive Bayes model bevezetése, illetve Support Vector Machine klasszifikátor bevezetése

Kötelező olvasmány:

- Lantz - Machine Learning with R, ch4

Hét 7 - Hasonlóság és Klaszteranalízis

Szövegek és dokumentumok hasonlósága/távolságának mérése, valamint különbözö klaszteranalizis technikák alkalmazása (hierarchikus illetve k-means klaszterezés)

Kötelező olvasmány:

- Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1), 43-48.

Hét 8 - Szövegek skálázása

Supervised és unsupervised módszerek használata arra, hogy egy vagy több dimenzió mentén jellemezzünk bizonyos szövegeket (pl.: politikai jobboldal - baloldal skála)

Kötelező olvasmány:

- Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American political science review*, 97(2), 311-331.
- Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science*, 52(3):705–722.

Hét 9 - Topic models

Unsupervised klasszifikáció, ami a dokumentumok témánkénti besorolását teszi lehetővé.

Kötelező olvasmány:

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Hét 10 - Prezentációk

Hét 11 - Eredeti adatgyüjtés: Webscraping, social media API

Bevezetés a web scraping technikákba, ami lehetővé teszi nagy mennyiségű online szöveg feldolgozását és letöltését.

Kötelező olvasmány:

- TBA

Hét 12 - Minden ami kimaradt vagy nem maradt rá idő

Témák közkívánatra illetve buffer.

Kötelező olvasmány:

- TBA