Kvantitatív szövegelemzés (MTA Poltext LAB)

SZTT-tipp

A kvantitatív szövegelemzés kurzus és oktatója, Máté Ákos nem először gazdagítja az ismeretbővítőink sorát: a módszertani fókuszú és gyakorlatorientált kurzus során az R statisztikai programcsomagot fogjátok használni, de előképzettség nem szükséges, ennek elsajátítására lesz szentelve az első pár alkalom. Ajánljuk politológusoknak, szociológusoknak, közgazdászoknak, de igazából bárkinek, aki kvantitatív szövegelemzéseket szeretne végezni bármilyen témában és bármilyen tudományterületen. Idén a kurzus a SZISZ és a poltextLAB kooperációjával valósul meg: a poltextLAB az MTA kvantitatív és Big Data módszerekkel szövegelemzéseket és szövegbányászatot végző inkubátor projektje, ennek a projektnek a vezető kutatói alkotják az oktatói gárdát: Máté Ákos, Ring Orsolya és Sebők Miklós, valamint a poltextLAB fiatal kutatója és egyben végzettünk: Székely Anna!

Követelmények: Félév végén 17 000 karakteres beszámoló + kód, félév közben rendszeres készülés az órákra + négy házi beadandó

SZTT-ből kit keress a kurzussal kapcsolatban: Viet, Bandi

„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 két ségtelenül előnyt jelent). Mivel a hangsúly az elméletek és módszerek alkalmazásán van, ezért különösebb matematikai/statisztikai előtanulmányok sem szükségesek.” (Máté, 2019)

Tematika

1. Kvantitativ Szövegelemzés alapjai.

Bevezető alkalom, alapfogalmak tisztázása, a terület áttekintése

• 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.

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.

• 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)

3. Szöveg mint adat 1.

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

• 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.

4. Szöveg mint adat 2.

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

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

5. Szótár alapú módszerek

Szentiment elemzés, szótárak konstrukciója és limitációi.

• 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.

6. Machine learning applikációk szövegek klasszifikációjára.

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

• Lantz - Machine Learning with R, ch4

7. Hasonlóság és Klaszteranalízis

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

• 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.

8. Szövegek skálázása

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

• 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.

9. Topic models

Unsupervised klasszifikáció, ami a dokumentumok témánkénti besorolását teszi lehetővé.

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

10. Prezentációk

11. Eredeti adatgyüjtés: Webscraping, social media API

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

12. Minden ami kimaradt vagy nem maradt rá idő

Témák közkívánatra illetve buffer.