Szövegbányászati és stilometriai eszközök

Ebben a dokumentumban azokat az eszközöket listázzuk és ismertetjük, amelyek valamilyen szövegbányászati lépéseket, illetve stilometriai elemzéseket végeznek. Ide soroltam minden olyan eszközt, ami nem nyelvi elemzést vagy kifejezetten csak előfeldolgozást végez.

Letölthető és/vagy beépíthető eszközök

stylo

- website: https://cran.r-project.org/web/packages/stylo/index.html
- github repo: https://github.com/computationalstylistics/stylo
- főbb függvények:
    - stylo() – szövegek közötti “távolságot” mér a leggyakoribb szavak alapján (paraméterezhető)
    - stylo.network() – az előbbi kiterjesztett változata, konszenzusfa ábrázolását teszi lehetővé
    - classify() – gépitanulás-alapú osztályozó algoritmusok (5 féle) futtatását teszi lehetővé; először be kell tanítani egy osztályozót egy megfelelően strukturált tanítóminta-készlettel, majd a vizsgálandó szövegeken futtatható az osztályozó
    - rolling.classify() – “görgetett” klasszifikáció, mintákra bontva vizsgálja a szöveget
    - oppose() – két szöveget vagy szövegcsoportot azonos méretű mintákra vág, és úgy hajtja végre a sylo()-nak vagy classify()-nak megfelelő összehasonlítást
    - imposters() – a General Imposters nevű szerzőazonosító módszert alkalmazza; összehasonlít egy szöveget a) olyanokkal, amelyeknek írói a potenciális szerzők között vannak és b) olyanokkal, amelyek íróinak kizárt a szerzősége
- a főbb függvények sok kis függvényt aggregálnak, és nagyobb elemzéseket tesznek lehetővé, de emellett még számos függvény (és ez a sok kis függvény önmagában is) létezik
- az összes függvény itt van leírva: https://cran.r-project.org/web/packages/stylo/stylo.pdf
- GPL licenc alatt elérhető, letölthető, felhasználható

rOpenSci

- https://ropensci.org/
- R-ben megírt, adatokkal való tudományos kutatáshoz használható, szabadon felhasználható csomagok

- kevés, de talán van köztük számunkra is használható, illetve ha mi csinálunk valamit, akkor azt be lehet ide nyújtani, hogy más is használhassa

MALLET

- MAchine Learning for LanguagE Toolkit
- http://mallet.cs.umass.edu/
- klasszikus statisztikai NLP csomag, ami a szokásos NLP lépéseken felül tud dokumentumklasszifikációt, témamodellezést és klaszterezést is
- tanítani kell saját anyagon
- az angolon kívüli egyéb nyelvekről nem esik szó, de mivel bármiylen adaton lehet tanítani, ezért, felteszem, nyelvfüggetlen --> mindenesetre sokat kell vele babrálni, miközben ugyanezek a funkciók rendelkezésre állnak más eszközökben, amikkel nem kell ennyit babrálni (ez a sztenderd nyelvfeldolgozó lépésekre vonatkozik)
- letölthető, kutatási célokra használható

GenSim

- NLP csomag, főleg szóvektoros módszerek gyűjteménye
- többféle témamodellezési módszer elérhető benne: LDA, NMF

WebSty

- https://ws.clarin-pl.eu/websty.shtml?en
- webes grafikus felület, amin keresztül kényelmesen lehet használni stilometriai eszközöket, főleg lengyelre
- API-n keresztül építhető be egy másik rendszerbe
- számos bemeneti formátumot elfogad, amit átalakít sima szöveggé -- > ehhez az Apache Tikát használja
- ami magyarra is működhet: többnyelvű szöveghasonlóság-elemzés
    - a magyar elemzéshez a UDPipe-ot használja
    - a UDPipe tsv-t bocsát ki, de a végső kimenet lehet XML is, vagyis nekik is meg kellett oldani a tsv2xml konverziót
- a szöveghasonlóság-elemzés 4 módszert takar:
    - szerzőségi elemzés: a korpusz szövegein elvégzett stilometriai elemzéseken alapulva --> a szövegek csoportosítása a stilometriai hasonlóság alapján
    - grammatikai elemzés: a szövegek csoportosítása a szöveg grammatikai jellemzői alapján
    - tartalomhasonlóság: a szövegek csoportosítása szó-együttelőfordulásokon alapulva
    - klasszikus szerzőségi elemzés: a szerzőség megállapítása szó-együttelőfordulásokon alapulva
- számos opciót lehet állítani, hogy milyen jellemzőket vegyen figyelembe az elemzéskor a rendszer
- teljeskörű elemzést nyújt, aztán abból a felhasználó választhatja ki, hogy mi érdekli

Voyant Tools

- https://voyant-tools.org/
- Létezik standalone változata is, VoyantServer a neve, Javában íródott.
- Teljesen nyílt forráskódú, kutatási célokra szabadon felhasználható.
- Amiket tud:
    - Szógyakorisági lista
    - KWIC konkordancia
    - A szógyakorisági adatok vizualizációja (szófelhő, grafikonok)
    - Kollokációk
    - A keresés maszkolható
    - Az eredmények exportálhatók (JSON, tsv, HTML)

Shtylo

- cikk: http://acta.bibl.u-szeged.hu/59065/1/msznykonf_014_423-436.pdf
- ezen a webes felületen keresztül kéne működnie?, de nem működik
- a stylo csomaghoz készült webes felület, ami az R nyelvben nem jártas DH-kutatóknak segít, plusz kiegészíti korpuszkezelési funkciókkal
- nem a magyarra specifikálták, hanem fekete doboz jellegű gépi tanulási módszerekkel próbálták megkerülni a problémát
- a Shtylo szerveren futtatja az R-t, webes felületet kínál a paraméterezéshez, adatbázisban tárolja az adatokat, és korpuszkezelő eszközkészletet is biztosít
- azért Shtylo, mert Sh(iny) + (s)tylo → a Shinyt kifejezetten az R szerveroldali futtatására tervezték
- github repó: https://github.com/dobijan/shtylo --> innen elvileg letölthető és újrahúzható az egész

Csak webes felülettel rendelkező, nem letölthető, nem beépíthető eszközök

TANIT

- TANIT = Text ANalysIs Tools
- dokumentumok összehasonlító elemzésére
- a magyarlánc kimenetén alapul, a tokenizálót, mondatrabontót és morfológiai egyértelműsítőt használja
- kifejezetten digitális bölcsészeti felhasználásra
- http://dighum.bibl.u-szeged.hu/tanit/
- a felhasználó az összehasonlító statisztikákat webes felületen böngészheti és táblázatos formában letöltheti további elemzés céljára
- egyszerű statisztikák:
    - mondatok száma
    - szavak száma (token)
    - átlagos mondathossz (szó/mondat)

    - különböző szóalakok száma (type)
    - különböző szótövek száma (lemma type)
    - minden egyes fő szófaji kódra az adott morfológiai elemzéssel ellátott szavak száma
- a művek szókincsgazdagságának jellemzésére használt metrikák:
    - Guiraud's R
    - Herdan's C
    - Text-Type Ratio (TTR)
    - CTTR
    - Dugast's U
    - Summer's index
- témamodellezés: Latent Dirichlet Allocation (LDA) (Blei et al., 2003)
    - Az LDA egy nagy dokumentumhalmazból képes előre megadott számú téma automatikus azonosítására. A látens témákat leírhatjuk azon szavak listájával, amelyeknek a kérdéses témában való szereplésének a feltételes valószínűsége a legnagyobb. Továbbá az LDA minden, a modellező halmazban szereplő és abban nem szereplő dokumentumhoz a témák egy eloszlását rendeli. Különböző művek ezen eloszlásainak összevetése érdekes tartalmi különbségekre világíthat rá.
    - két lehetőség:
        - előre betanított modell alkalmazása egy feltöltött dokumentumon: ez jelenleg az MNSZ2 szépirodalmi alkorpuszán van tanítva
        - a feltöltött dokumentumokon tanítunk egy modellt: ehhez általában kevés az adat, ezért a feltöltött dokumentumokat tovább vagdossák kisebb dokumentumokra --> csúszóablakkkal állítható
    - a témák száma állítható
    - a MALLET LDA-ját használják
- csak böngészőből működik
- lehetséges előfeldolgozási lépések:
    - stopszavak listája
    - kisbetűsítés
    - számok eltávolítása
    - URL-ek eltávolítása
- a feltöltött dokumentumokat az elemzés elkészülte után azonnal törlik a szerverről
- az eredményt megjelenítik, és xslx formátumban letölthetővé is teszik
- az eredményfájlt 24 óráig tárolják a szerveren, utána törlik

Interactive Text Mining Suite

- https://languagevariationsuite.shinyapps.io/TextMining/
- R & Shiny
- elsősorban adatvizualizációs eszköz kifejezetten DH célokra
- PDF vagy TXT fájlt kell feltölteni, a PDF-nél azzal a kikötéssel, hogy ha nem konvertálja és javítja a felhasználó magának előre a szöveget, akkor az OCR-kimeneten nem fog olyan jó eredményeket kapni
- metaadatokat is fel lehet tölteni megadott formátumokban
- lehet előfeldolgozási lépéseket is csináltatni:

    - punktuáció eltávolítása (kivételekkel)
    - kisbetűsítés
    - számok eltávolítása
    - hivatkozások eltávolítása
    - URL-ek eltávolítása
    - HTML-elemek eltávolítása
    - stopszavak eltávolítása, akár saját listával
    - tövezés (van magyarra is)
- vizualizációs eszközök és mögötte levő elemzések:
    - raw frequency table
    - szófelhők
    - mondat- és szóhossz
    - KWIC állítható kontextusban
    - punktuációelemzés (Adam J. Calhoun elemzésén alapulva)
    - klaszteranalízis számos beállítással
    - témamodellezés (LDA, STM, metaadatelemzés) számos beállítással
- csak webes felület, nem letölthető, nem beépíthető

AVOBMAT

- AVOBMAT: Analysis and Visualization of Bibliographic Metadata and Texts
- leegyszerűsített (többek között feltöltés, előfeldolgozás és lexikai gazdagság elemző nélküli) béta verzió kipróbálható a COVID-19-es adatbázison.
- AVOBMAT-ot bemutató cikk, elemző és egyéb funkciók listája, valamint használati segédanyag (COVID-19-es) példákkal
- bibliográfiai adatokat és korpuszokat is lehet vele elemezni
- elemzés, vizualizáció és export funkciók
- webes alkalmazás
- a bibliográfiai adatok feldolgozásához Zotero kollekciókat lehet feltölteni csv és rdf formátumban
- a szövegek feltölthetők txt, pdf, doc(x) és xml formátumokban, amikből az Apache Tika python könyvtárral nyerik ki a szöveget
- előfeldolgozó lépések:
    - kisbetűsítés
    - számok eltávolítása
    - nem alfabetikus tokenek eltávolítása
    - kötőjelek eltávolítása
- opcionális további lépések:
    - lemmatizálás
    - stopszó szűrés
    - punktuáció szűrés
- a nyelv specifikálása történhet kézzel, vagy lehet használni a langdetect eszközt
- a stopszavak és punktuációs jelek eltávolításához a spaCy-t használták
- a lemmatizáláshoz bizonyos nyelvek (így a magyar) esetében a LemmaGen eszközt használták, bizonyosakhoz meg a spaCy-t
- minden elemzés online történik
- a nyers és az előfeldolgozott szöveg is tárolva van

- a szerzők nemét automatikusan állapították meg a Python sexmachine csomagjával
- keresés: Elasticsearch --> strict, fuzzy and proximity search & regexes keresés
- mindent ki lehet exportálni a session végén
- a kiexportált kimenetek más eszközök által is használható formátumban vannak
- a session végén minden törlődik
- a tartalomelemzés elemei:
    - szó n-gram viewer:
        - évi lebontásban lehet megnézni a specifikált n-gramokat
        - max. 5gram
        - az n-gramok az előfeldolgozás során előállnak
        - normalizálva is lehet nézni: az n-gramok száma le van osztva az adott évben előforduló összes szó számával
    - szófelhő:
        - significant text cloud: a keresett szóhoz leginkább kapcsolódó szavakat adja ki
        - TagSpheres: a keresett szó kontextusát rajzolja ki
    - témamodellezés:
        - az egyes dokumentumok témáját adja ki, illetve a hasonló tematikájú dokumentumokat rakja össze
        - jsLDA: in-browser topic modelling --> kikalkulálja és grafikailag meg is jeleníti a témamodelleket
        - az eredményeket 6 különböző módon lehet kiexportálni: document topics, topic words, topic summaries, topic-topic connections, doc-topic graph file (for Gephi) and complete sampling state (docID, word and topic number)

Lexos

- http://lexos.wheatoncollege.edu/upload
- csak webes felület, letölteni nem lehet, de ötletet meríteni belőle igen
- feltöltés: fájlok, URL-ek alapján
- szokásos előfeldolgozási lépések
- vizualizáció:
    - szófelhő
    - multicloud
    - bubble viz
    - rolling window
- elemzések:
    - token- és karakterszintű statisztikák
    - dendrogram (fa reprezentálásra szolgáló eszköz, más néven klaszteranalízis) --> opciók:
        - távolságmetrika, ami az egyes dokumentumok egymástól való távolságát számolja
        - orientáció: balról jobbra vagy lentről fölfele
        - token- vagy karakterszintű n-gramok alapján, ahol az n is állítható
        - gyakoriságnormalizálás: abszolút, relatív, tf-idf
        - válogatás: gyakoriság alapján

    - k-means (klaszterezés)
    - konszenzusfa
    - hasonlósági keresés
    - gyakorisági listák
    - tartalomelemzés
- a webes felületen csak az eredmény egy része látható, ami jól megjeleníthető és elfér, a többi eredményt ki lehet exportálni: csv-ben vagy valamilyen képformátumban

Futottak még

ProQuest TDM Studio

- https://about.proquest.com/products-services/TDM-Studio.html
- fizetős, ingyen demólehetőséget lehet igényelni
- csak ProQuest kollekciókkal rendelkező könyvtáraknak és intézményeknek használható --> ProQuest: könyvtáraknak és kutatóknak biztosítanak technológiát tartalommenedzsmenthez

Signature

- http://www.philocomp.net/texts/signature.htm
- a legegyszerűbb stilometriai mutatókat lehet vele előállítani
- úgy tűnik, mintha csak vindózon lenne futtatható
- letölthető, kutatói célokra használható
- elég fapados felület

JGAAP

- https://github.com/evllabs/JGAAP
- szerzőazonosításra szolgáló eszköz grafikus felülettel --> a megadott lehetséges szerzők közül kiválasztja a legvalószínűbbet egy adott szöveghez
- Javában van írva
- elég hiányos a dokumentáció

Stylene

- http://www.stylene.be/
- készül egy újabb verzió, addig csak egy flamand(?) webes demó érhető el, illetve egy olvashatósági demó, ami megmondja egy kisebb szövegrészletre, hogy az mennyire olvasható

Gale Digital Scholar Lab

- https://www.gale.com/intl/primary-sources/digital-scholar-lab
- csak kérni lehet egy trialt, de amúgy semmilyen szinten nem hozzáférhető

- sok dokumentumot lehet feltölteni, keresni, kezelni
- NER, témamodellezés, szófaji elemzés
- a felhasználók adatai sokáig megmaradnak, projektekbe szervezhetik őket
- az elemzések kimenete megosztható