TANIT – Magyar nyelvű szövegeket elemző eszköz összehasonlító digitális bölcsészeti feladatokhoz

Labádi Gergely[1], Farkas Richárd[2], Nagy Roland[2], Péter Róbert[3]

[1] Szegedi Tudományegyetem,
Magyar Nyelvi és Irodalmi Intézet
[2] Szegedi Tudományegyetem,
Informatika Intézet
rfarkas@inf.u-szeged.hu
[3] Szegedi Tudományegyetem,
Angol-Amerikai Intézet
robert.peter@ieas-szeged.hu

Kivonat: Cikkünk a TANIT (Text ANalysIs Tools) rendszer célkitűzését, funkcióit és használatát mutatja be. A TANIT rendszer célja, hogy magyar nyelvű szövegek számítógépes nyelvészeti feldolgozásával dokumentumok összehasonlító elemzéséhez szükséges statisztikákat kigyűjtsön. Ez a webszolgáltatás létező nyelvtechnológiai elemzőlánc kimenetére épülő aggregált statisztikákat számít ki, témamodelleket épít és ezeket olyan formátumban adja át a digitális bölcsész felhasználónak aki utána programozói ismeretek nélkül is fel tudja ezt használni kutatásaiban.

1 Bevezetés

Mindannyian tudjuk és tapasztaljuk, az információs társadalom korszaka jelentős kihívás a humán tudományok számára: a változás egyaránt érinti a tudás társadalmi intézményeit, legitimációját, ám ugyanúgy kihat mindennapjainkra, tudományos gyakorlatunkra. Mindez alkalmat nyújt – valójában sürget – a humántudományi kutatások tárgyának, módszerének és közegének újragondolására [1]. Ez az egyik legfontosabb célja a digitális bölcsészetnek (*digital humanities*). A nyugati tudományosságban a digitális bölcsészet a 2000-es évek második felére szervezetileg, intézményileg végleg és egyértelműen áttört, azaz vannak képzések, folyóiratok, kutatóközpontok, konferenciák, a bölcsész állásoknál rendszerint elvárás valamiféle DH-képesség vagy -gyakorlat.

A digitális tudományos kutatás formabontó és úttörő kutatási lehetőségekkel kecsegtet a bölcsészettudományokban, mivel a szövegek és az ezekhez rendelt metaadatok elemzéséhez számtalan új lehetőséget kínálnak, melyeket e tudományterületen korábban nem alkalmaztak. Az új digitális módszerek és eszközök használatát a világszerte gomba módra szaporodó digitális bölcsészet tanszékeken oktatják. A hallgatók megtanulják például miként kell adatbázisokat létrehozni, szövegeket felcímkézni. Született már olyan egyetemi tankönyv is, amely kifejezetten a nyelv és irodalom szakosokat célozza meg a hatalmas szövegkorpuszok és

adatbázisok statisztikai elemzéséhez és ábrázolásához használt R programnyelv megismertetésével, amelynek van kifejezetten nyelvészeti és irodalmi elemzésekhez használható csomagja is [2].

A Szegedi Tudományegyetemen létrejött – jelentős részben a 2017. szeptemberében elhunyt Labádi Gergelynek köszönhetően – egy digitális bölcsészet fókuszú informális együttműködés különböző karok különböző intézetei közt: Magyar Nyelvi és Irodalmi Intézet (BTK), Angol-Amerikai Intézet (BTK), Informatika Intézet (TTIK), Klebelsberg Kuno Könyvtár.

Laptopos bemutatónkban demonstráljuk a TANIT (TANIT – Text ANalysIs Tools) rendszert, ami ennek az együttműködésnek az első materializálódott eredménye. A TANIT rendszer célja, hogy kis mennyiségű, tipikusan néhány tucat, magyar nyelvű dokumentum mélyebb számítógépes nyelvészeti feldolgozásával dokumentumok összehasonlító elemzéséhez szükséges statisztikákat kigyűjtsön. A TANIT egy publikusan elérhető webes felület (http://dighum.bibl.u-szeged.hu/tanit), ahol a bemeneti dokumentumok feltöltése után lehetőség van az elemzés néhány paraméterének beállítására, majd a felhasználó az összehasonlító statisztikákat webes felületen böngészheti és táblázatos formában letöltheti további elemzés céljára. A TANIT célfelhasználója az a digitális bölcsész, aki ismeri a számítógépes nyelvészetben rejlő lehetőségeket, de nem akar saját nyelvtechnológiai pipelinet építeni, csak a számára legfontosabb statisztikákat megkapni.

A TANIT funkcióinak tervezésekor néhány valós összehasonlító digitális bölcsészeti kutatás nyelvtechnológiai igényeiből indultunk ki. Például Labádi Gergely egyik összehasonlító irodalomtudományi tanulmányában [3] azt a kérdést igyekszik megválaszolni, hogy Jókai *A kőszívű ember fiai* című művének különböző átdolgozásai mennyire hasonlítanak az eredeti műhöz, mennyire tekinthetőek ugyanazon műnek. Például Nógrádi átdolgozása egy „modernizált" változat. Pedagógusok szerint ugyanis a mai diákok nem értik – és így nem is olvassák – a klasszikus (és kötelező) irodalmat, kifejezetten a regények nyelvezetét találják problémásnak, a nyelvre helyezik a hangsúlyt, ami érinti az igeidőket, a latin, német kifejezéseket, és természetesen magukat a mondatstruktúrákat is.

A kőszívű ember fiai különböző változatainak összehasonlító kutatásához olyan nyelvtechnológiai kérdések megválaszolására van szükség, mint

- Az egyes művekben milyen a szavak és mondatok hosszának eloszlása?
- Mennyire dinamikus/eseményorientált a mű? Ez az igék és egyéb szófajok arányával közelíthető.
- Az egyes témák ugyanolyan súllyal szerepelnek-e a változatokban?

2 Kapcsolódó munkák

Nemzetközi szinten számos tudományos publikáció [4,5,6,7] és elérhető digitális eszköz[1] született hasonló célzatú összehasonlító digitális bölcsészeti kutatások

[1] Mallet: http://mallet.cs.umass.edu
Interactive Text Mining Suite: https://languagevariationsuite.shinyapps.io/TextMining/

támogatására. Az elérhető eszközök általában az angol nyelvre koncentrálnak. Vannak köztük olyanok, amelyek akár tucat nyelv kezelését is ígérik, de ez általában egyszerű szótövezést takar csak. Megvizsgáltuk ezen eszközök pontosabb és mélyebb magyar nyelvre kidolgozott eszközökkel való kibővítésének lehetőségét, de technológiai oldalról ez nem volt kivitelezhető. Ezen vizsgálódások folyományaként döntöttünk úgy, hogy saját rendszert implementálunk. A TANIT tervezése során azonban erősen építettünk a megismert nemzetközi megoldásokban alkalmazott elemzési lépésekre.

Magyar nyelvtechnológiai oldalról tekintve is több kapcsolódó tudományos publikáció és elérhető eszköz létezik (például e-magyar[2] [8], magyarlanc[3] [9], hun* eszközök[4] [10]). Az elérhető magyar nyelvű szövegek számítógépes nyelvészeti feldolgozását támogató elemzők és elemzőláncok kimenetele már strukturált metaadat, például egy mondat szavainak morfológiai leírása, de még nem kezelhető könnyen, programozói ismeretek szükségesek a kiment feldolgozásához, ugyanis ezekből még aggregált statisztikák leszámolására van szükség egy összehasonlító digitális bölcsészeti kutatáshoz. Továbbá az összehasonlítások talán legfontosabb szempontja a dokumentumok tartalmi elemzése és legjobb tudomásunk szerint nem érhető el szabadon felhasználható témamodellező (topic modelling) eszköz magyar nyelvű szövegek elemzésére. A TANIT célja, hogy létező nyelvtechnológiai elemzőlánc kimenetére épülő aggregált statisztikákat kiszámítsa, témamodelleket építsen és ezeket olyan formátumban adja át a digitális bölcsész felhasználónak aki utána programozói ismeretek nélkül is fel tudja ezt használni kutatásaiban.

3 Alapstatisztikák

A feltöltött dokumentumokon a TANIT mondat és szószegmentációt, majd szófaji egyértelműsítést futtat. Egyéb konfigurálható előfeldolgozási lépések után a TANIT minden dokumentumra külön leszámolja majd egyetlen táblázatba szervezve visszaadja az alábbi egyszerű statisztikákat:

- 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

Az alapstatisztikák mellett a művek szókincsgazdagságának jellemzésére két statisztikát számol a TANIT. A magyar irodalomtudományi szakirodalomban elfogadott Guiraud-index:

Stylo: https://sites.google.com/site/computationalstylistics/

[2] http://e-magyar.hu/

[3] http://www.inf.u-szeged.hu/rgai/magyarlanc

[4] http://mokk.bme.hu/en/eszkozok/

$$\frac{type}{\sqrt{token}}$$

A Guiraud-index nem alkalmas a szöveghosszból fakadó különbségek semlegesítésére. Ennek feloldására javasolt Herdané [11] egy másik képletet:

$$\frac{\log type}{\log token}$$

4 Témamodellezés

Az összehasonlító elemzések tartalmi szintjéhez a TANIT két különböző eszközzel járul hozzá. Mindkét esetben a Latent Dirichlet Allocation (LDA) valószínűségi témamodellező (topic modelling) algoritmust [12] alkalmazzuk. Az LDA egy nagy dokumentumhalmazból képes előre megadott számú téma automatikus azonosítására. A látens témákat leírhatjuk azon szavak listájával amelyek a kérdéses témában szereplésének feltételes valószínűségei a legnagyobbak. Továbbá az LDA minden, a modellező halmazban szereplő és abban nem szereplő dokumentumhoz a topikok egy eloszlását rendeli. Különböző művek ezen eloszlásainak összevetése érdekes tartalmi különbségekre világíthat rá.

A TANIT két különböző megközelítése témamodellezésre abban tér el, hogy milyen dokumentumhalmazon modellezi a témákat. A témamodell alkalmazása (dokumentumok topik eloszlásának számítása) már ugyanúgy működik:

- Lehetőségünk van egy előre betanított LDA használatára a feltöltött dokumentumainkon. Ehhez jelenleg egy az MNSz2 [13] szépirodalmi alkorpuszán tanított modellt biztosítunk. Folyamatban van a Magyar Elektronikus Könyvtár (MEK) feldolgozása, ami eggyel nagyobb nagyságrendű szöveghalmazt biztosít majd. Ezen a korpuszon már érdemes lehet különböző szempontok szerint szűrt részkorpuszokon – például 19. sz. magyar irodalom – különböző LDA modelleket tanítani.
- Vagy magán a feltöltött dokumentumokon is taníthatunk egy LDA modellt. A TANIT kevés számú mű elemzését célozza meg, ami nem elégséges mennyiség az LDA tanításához. Ezért lehetőség van egy csúszó ablakkal a művek kisebb kvázi-dokumentumokra bontására majd ezeken a kvázi-dokumentumokon LDA tanítására. Azt tapasztaltuk, hogy néhány száz szó méretű ablak már elégséges együtt előfordulási kontextust biztosít az LDA számára, de az ablakméretet a felhasználó is állíthatja. Ennek a megközelítésnek az az előnye, hogy a témák kizárólag a feltöltött műveket bontják le témákra. A jövőben tervezzük a csúszó ablakokból a mű cselekményének időbeli fejlődésének vizualizációját és dokumentumok közti összevetését is.

A második megközelítés alkalmazásával Labádi [3] tudta azonosítani hogy Nógrádi átdolgozott, modernizált verziója “a Baradlay-gyerekek mellé az édesanyjukat érintő

eseményszálakat domborította ki", míg az eredeti Jókai "a család köré fonódó intrikát emeli ki".

5 Technológiai megvalósítás

A TANIT egy böngészőből elérhető szolgáltatás. A vékony front-end egy JSP weboldal amin keresztül feltölthetjük a szerverre az elemzés alanyául szolgáló dokumentumokat. A dokumentumokat egyesével megadva vagy egy .zip állományba csomagolva is feltölthetjük. Nyers szövegfájlok (.txt) feldolgozására van a rendszer jelenleg felkészítve aminek karakterkódolását előre beállíthatja a felhasználó.

A felhasználó az alábbiakat konfigurálhatja a weboldalon ha az alapértelmezett beállításokkal nem elégedett:

- saját stopword szótár használata (feltöltése)
- kiegészítő szótőtár hozzáadása
- kisbetűsítés, számok cseréje
- LDA-nál:
    - előre betanított modellek közül választás
    - ha saját belső modellt tanul, akkor a csúszóablak mérete
    - topikok száma

Az elemzés elindításával a dokumentumok feltöltésre kerülnek egy egyetemi szerverre. A dokumentumokat a nyelvtechnológiai elemzések elkészülte után azonnal töröljük a szerverről. Az egyidejű felhasználók kiszolgálása session alapú kiszolgálása biztosítja, hogy azok nem férhetnek hozzá mások adatához.

A server oldal egy Tomat web application server, ami Java nyelvű kódot futtat. A nyelvtechnológiai elemzésekhez a magyarlanc [9] szó- és mondatszegmentációját valamit szófaji egyértelműsítését (ami a környezetfüggő szótöveket is megadja) használtuk fel. A saját stopword listák és kiegészítő szótőtárak miatt kisebb módosításokat kellett végrehajtanunk a magyarlancon. Az LDA témamodellezést a Mallet[5] [14] könyvtárainak felhasználásával végezzük. A leszámolt statisztikákat visszaadjuk a front-endnek megjelenítésre és egy .xlsx fájlt is megkonstruálunk belőlük. Ebben a fájlban különböző lapokra (spreadsheet) kerülnek a különböző elemzések, az egy lapon szereplő táblázatok sorai pedig az egyes dokumentumoknak felelnek meg. Az .xlsx fájlt a felhasználó letöltheti és azon könnyen további aggregált statisztikákat számolhat. Az eredmény fájlt 24 óráig tároljuk a szerveren, utána törlésre kerül.

6 Összegzés és további tervek

Laptopos bemutatónkban bemutatjuk a http://dighum.bibl.u-szeged.hu/tanit rendszert, ami összehasonlító digitális bölcsészeti kutatások támogatására készült

[5] http://mallet.cs.umass.edu

nyelvtechnológiai rendszer. A rendszer célja, hogy bármiféle programozói vagy nyelvtechnológiai ismeret nélkül használhassák azt bölcsész kutatók és hallgatók.

A TANIT első verziója elkészült és publikusan elérhető. A jövőben tervezzük funkcionalitásának bővítését, például különböző témamodellek biztosításával, szintaktikai elemzőkre építve grammatikai hasonlósági mértékek alkalmazása művek közt, plágiumkeresők implementálása stb.

Köszönetnyilvánítás

A munka az EFOP-3.6.1-16-2016-00008 azonosítójú, EU társfinanszírozású projekt támogatásával készült el. Farkas Richárd kutatásait az MTA Bolyai János ösztöndíja támogatta.

Bibliográfia

1. http://digibolcsesz.ek.szte.hu/
2. Jockers, M. L. (2014): Text Analysis with R for Students of Literature (Quantitative Methods in the Humanities and Social Sciences). Springer-Verlag, Cham (2014);
3. Labádi, G.: Latra vetett szavak: A kőszívű ember fiai kiadásairól és átiratairól. Digitális Bölcsészet (2018) in press.
   online változat: http://www.academia.edu/34473500/Labadi_Latravetettszavak.pdf
4. Ramsay, S.: Reading Machines: Toward an Algorithmic Criticism University of Illinois Press, Champaign, IL (2011)
5. Jockers, M. L.: Macroanalysis: Digital Methods and Literary History. University of Illinois Press, Urbana (2013)
6. Moretti, F. Distant Reading. Verso, London (2013)
7. Schreibman, S., Siemens, R. and Unsworth, J. (eds) A New Companion to Digital Humanities. Blackwell, Oxford (2016)
8. Váradi T., Simon E., Sass B., Gerőcs M., Mittelholcz I., Novák A., Indig B., Prószéky G., Farkas R., Vincze V. Az e-magyar digitális nyelvfeldolgozó rendszer. Magyar Számítógépes Nyelvészeti Konferencia (2017)
9. Zsibrita, J., Vincze,V., Farkas,R. magyarlanc: a toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP 2013, pp. 763–771 (2013)
10. Halácsy P., Kornai A., Oravecz Cs. Hunpos an open source trigram tagger. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (2007)
11. Herdan, G. The Advanced Theory of Language as Choice and Chance. Springer-Verlag, Berlin, Heidelberg, New York. (1966), 75–77.
12. Blei, D., Ng A., Jordan, M. Latent dirichlet allocation. Journal of Machine Learning Research 3 (2003), 993-1022.
13. Oravecz Cs., Váradi T., Sass B.: The Hungarian Gigaword Corpus. In: Proceedings of LREC (2014)
14. McCallum, A. MALLET: A Machine Learning for Language Toolkit. (2002)