**Nagy Gyula**
SZTE Klebelsberg Könyvtár
SZTE Neveléstudományi Doktori Iskola

Networkshop 2016 – Debrecen, 2016. április 1.

**Tudománymetriai és tartalmi elemzések szövegbányászati módszerekkel**

Hamarosan elérjük azt az állapotot – a digitális szingularitást – amikor minden valaha létező és folyamatosan keletkező (analóg és már digitális formában született) információs anyag (szöveg, kép, hang, mozgókép) elérhető lesz a hálózaton (*Nagy*, 2014). A teljes szövegű adatbázisok már egyébként is jó néhány éve megjelentek az életünkben. A kemény tudományok korábban elkezdték kihasználni az általuk biztosított előnyöket, de mára a társadalomtudományok és a bölcsészettudományok is erősen építenek a teljes szöveg által nyújtott előnyökre. Egyre több és több tudományág próbálja kihasználni a teljes szövegű korpuszok lehetőségeit: a digitális bölcsészet, a digitális filológia, számítógépes nyelvészet, oktatási adatbányászat, tudománymetria, számítógépes szociológia (*Holl, 2015).*

A jelenség közös alapját az adat- és szövegbányászat adja. A szövegbányászat az adatbányászatból alakult ki, ezért mindenképpen szükséges a két terület viszonyának a tisztázása. A legfőbb különbség a két terület között, hogy míg az adatbányászat (data mining) strukturált adatokkal dolgozik, (amelyek sok esetben numerikusak), addig a szövegbányászat (text mining) strukturálatlan szövegeket használ input alapanyagként, ezért tehát elmondhatjuk, hogy a szövegbányászat az informatika szöveges dokumentumok elemzésével foglalkozó részterülete (*Tikk,* 2007).

A szövegbányászatra szöveges adatbányászatként (text data mining) is szoktak hivatkozni, illetve elterjedt még a szöveg analízis (text analytics) terminus is. Kialakulása az üzleti intelligenciához (business intelligence) köthető és az 1950-es évek végére tehető, az IBM egyik lapjában így ír Luhn: „Az adatfeldolgozó gépek használhatóak dokumentumok auto-referálására és auto-kódolására.” (*Luhn*, 1958). A számítógépes infrastruktúra, a mesterséges intelligenciakutatás és a gépi tanulóalgoritmusok fejlődésének hatására az 1990-es években kezdtek el újra komolyabban foglalkozni a szövegbányászattal, az adatbányászat egy speciális aleseteként. Például a szegedi kutatók már az 1990-es évek vége óta jelen vannak a területen, „Szeged Treebank” néven építettek korpuszgyűjteményt, melyet számtalan szövegbányászati és számítógépes nyelvészeti kutatás felhasznált (*Csendes, Csirik, Gyimóthy,* és *Kocsor*, 2005).

Az eljárás lényege abban áll, hogy nagymennyiségű strukturálatlan szöveg automatikus gépi elemzésével próbálunk olyan következtetéseket levonni a szöveget illetően, amelyek esetében

akár az is elképzelhető, hogy explicit módon nincsenek benne a szövegben, vagy csak rejtetten, esetleg az óriási mennyiségű szövegkorpuszokban elvesznek ezek az egyébként lényeginek tekinthető információk. A szövegbányászat különböző lépések, eljárások, algoritmusok felvonultatásával rejtett mintázatokat próbál találni a szövegekben, amelyekből azután a módszert alkalmazó hozzáértő kutató különféle tudományos következtetéseket tud levonni (*Tikk*, 2007).

A módszer számtalan alkalmazási területét lehetetlen bemutatni, azonban néhány példát mégis kiragadnánk. Számos elemzés származik az orvosbiológia (*Cohen*, és *Hersh*, 2005; *Vincze, Szarvas, Farkas, Móra* és *Csirik*, 2008), az üzleti informatika (*Ghosh, Haider* és *Sen*, 2015), a számítógépes nyelvészet (*Schneider*, 2014), vagy a könyvtártudomány (*Nagarkar*, 2015) területéről, de szinte minden olyan diszciplína esetén lehet példát találni használatára, ahol nagyobb mennyiségű szövegek fordulnak elő.

A hivatkozásvizsgálatok és a szövegbányászatot segítségül hívó tartalmi elemzések sok más, az informatika által nyújtott megoldással együtt hamarosan be fognak épülni minden tudományág eszköztárába. Azonban ma még nincsenek jól megalapozott és széles körben használható sztenderdjeink, vagy univerzális, használatra kész szoftvercsomagjaink, ezért új utakat kell keresnünk és kísérleteznünk kell: ez egy szükségszerűség, mivel az ígért állapot nem jön el magától, az információs szakembereknek kell kitalálnia, megalkotnia és szolgáltatnia ezeket a megoldásokat a tudósok számára. Az előadás egy ilyen kísérletezésről számol be, a hivatkozásvizsgálatok és a szövegbányászati elemzések területéről.

A kutatásban kiválasztásra került egy elismert folyóirat, amely az egyik legfontosabb és legrégebbi magyar neveléstudományi folyóirat, a Magyar Pedagógia. 1892-ben alapították és még napjainkban is megjelenik. Az összes évfolyam digitalizálásra került és ezzel létrejött egy 50.000 oldalas szövegkorpusz a folyóiratcikkekből (kb. 6.500 tanulmány), továbbá az összes metaadat bekerült egy jól strukturált adatbázisba, amelyet katalogizáló könyvtárosok építettek. A teljes szövegű korpusz szövegbányászattal került elemzésre: ez a megoldás a rejtett információkat és kapcsolatokat tehet láthatóvá a szövegstruktúrából. Például láthatjuk, hogy mennyi cikk foglalkozik az oktatási reformmal és hogyan változik ez a téma, ahogy az idő telik. A szövegbányászati vizsgálódással párhuzamosan egy tudománymetriai elemzés is végrehajtásra került, mivel a teljes szöveg lehetővé tette ezt. A kutatás megcélzott kimenetei között szerepelt egy hatalmas gráf megalkotása, az összes szerzővel és hivatkozással.

A kutatás egy próbaprojekt a társadalom- és a humántudományok folyóiratainak tudománymetriai és szövegbányászati elemzésére, mivel ezek jóval kevésbé reprezentálódnak a tudományos hivatkozásokat feltáró adatbázisokban. (*Nagy*, 2016) A fő cél az volt, hogy a

kutatás nyomán új eszközök kerülhessenek bevezetésre az akadémiai szférában dolgozó könyvtárosok eszköztárába, amelyek sok lényeges eredményt produkálhatnak a jövőben a különféle tudományágakban az információs szakemberek segítségével.

Irodalomjegyzék:


1. Cohen, A. M., Hersh, W. R. (2005): A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6. 1. sz. 57-71.
2. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A. (2005): The szeged treebank. In: Text, Speech and Dialogue. Springer, Berlin Heidelberg. 123-131.
3. Ghosh, R., Haider, S., Sen, S. (2015): An integrated approach to deploy data warehouse in business intelligence environment. Computer, Communication, Control and Information Technology 1-4.
4. Holl András (2015): Szövegbányászat, adatbányászat, ismeretfeltárás. Új lehetőségek a tudományos kommunikációban Magyar Tudomány, 176. 6. sz. 680-685.
5. Luhn, H. P. (1958): A business intelligence system. IBM Journal of Research and Development, 2. 4. sz. 314-319.
6. Nagarkar, S. P. (2015): Text mining: an analysis of research published under the subject category ‘Information Science Library Science’ in Web of Science Database during 1999-2013. Library Review, 64. 3. sz.
7. Nagy Gyula (2014): Megy-e a digitalizálás által a világ elébb? Avagy mi végre digitalizálunk? Információs Társadalom, 14. 3. sz. 44-52.
8. Nagy Gyula (2016): Tudománymetria és neveléstudomány. Iskolakultúra, 26. 2. sz. 50–62.
9. Schneider, G. (2014): Applying Computational Linguistics and Language Models: From Descriptive Linguistics to Text Mining and Psycholinguistics. Department of English Institute of Computational Linguistics University of Zurich.
10. Tikk Domonkos (2007, szerk): Szövegbányászat. Typotex, Budapest.
11. Vincze, V., Szarvas, G., Farkas, R., Móra, G., Csirik, J. (2008): The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics, 9. Suppl 11. S9.