

**AZ OKTATÁS, A KUTATÁS ÉS  
A KÖZGYŰJTEMÉNYEK DIGITÁLIS  
TRANSZFORMÁCIÓJA FELSŐFOKON**

**NETWORKSHOP 2024  
33. Országos Informatikai Konferencia**

**2024. április 3–5.  
Eszterházy Károly Katolikus Egyetem, Eger**

# AZ OKTATÁS, A KUTATÁS ÉS A KÖZGYŰJTEMÉNYEK DIGITÁLIS TRANSZFORMÁCIÓJA FELSŐFOKON

NETWORKSHOP 2024  
33. Országos Informatikai Konferencia

2024. április 3–5.  
Eszterházy Károly Katolikus Egyetem, Eger

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület  
Budapest, 2024



**HUN-REN**  
Magyar Kutatási Hálózat

# NETWORKSHOP

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória  
Korrektúra: Danyi Melinda  
Angol nyelvi lektor: Cseresnyés Dorottya

Networkshop 2024 konferencia előadásainak közleményei  
Eszterházy Károly Egyetem, Eger  
2024. április 3–5.

ISBN 978-615-82243-2-1  
DOI: <https://doi.org/10.31915/NWS.2024>

Kiadja a HUNGARNET Egyesület  
az MTA Könyvtár és Információs Központ közreműködésével  
Budapest  
2024

Borítókép: [freepik.com](https://www.freepik.com)

## TARTALOMJEGYZÉK

<b>Előszó</b> .....	5
<b>Ungváry Rudolf</b> A MARC21 formátum kettős szerkezete és a formátum felhasználói szintjének fordításai .....	7
<b>Holl András, Andódy Katalin</b> Adatbányászati gyakorlatok repositóriumra és MTMT-re.....	16
<b>Simon András</b> Mesterséges intelligenciával támogatott adatgazdagítás a Nemzeti Levéltárban.....	22
<b>Soós Gábor, Rövid András, Ormos Pál</b> V2X – A járművek közötti kommunikáció kihívásai .....	29
<b>Csernai Zoltán</b> Egy online tanulást támogató portál kurzusának vizsgálata Big Data adatelemző módszerekkel.....	36
<b>T. Nagy László, Németh Áron</b> A mesterséges intelligencia (MI) teológiai kompetenciái .....	45
<b>Mészáros Erika</b> Kodaktól a jövőig – Egy könyvtári digitalizálás szinterei.....	55
<b>Frankó Máté, Sándor Ákos</b> Adatvizualizáció a könyvtári menedzsmentben: fejlesztések az SZTE Klebelsberg Könyvtár döntéstámogató rendszerében.....	63
<b>Hernek István</b> A felhasználóképzés szintjei az SZTE Klebelsberg Könyvtárban: az elsőévesektől a kutatókig .....	72
<b>Némethi-Takács Margit, Borbély Mária</b> Bibliográfiai kapcsolatok az általános megjegyzés adatmezőben.....	78
<b>Dobás Kata, Tüskés Anna</b> A magyar irodalomtörténet bibliográfiájának migrációja az ITIdata szemantikus adatbázisba .....	87
<b>Horváth Péter</b> A kanonikus magyar költészet versformakeresője.....	96
<b>Sebestyén Ádám, Sárközi-Lindner Zsófia</b> Történeti források szemantikus feldolgozása – Az ELTEdata adatbázis új gyűjteményei .....	105

<b>Bolya Mátyás</b>	
Lyukkártya és népdalrendezés – Egy mechanikus népzenei adatbázis digitális rekonstrukciójának lehetőségei.....	112
<b>Kovácsházy Tamás</b>	
Az idő, mint alapvető infrastruktúra, az idő szerepe az adatközpontban.....	121
<b>Albert Ágota Katalin</b>	
A mesterséges intelligencia használatának követelményei az oktatási szektorban, különös tekintettel a mesterséges intelligencia használatáról szóló rendeleltre.....	129
<b>Varga Emese</b>	
Digitális szövegszerkesztés a dHUpla keretrendszerében .....	135
<b>Nemoda Zsuzsanna, Héjja Balázs, Nagy Andor, Tóth Máté</b>	
A Pest Megyei Digitális Könyvtár fejlesztése .....	141
<b>Nagy Dóra, Sándor Ákos</b>	
Voice2text: a hanganyagátírás lehetőségei MI segítségével.....	149
<b>Kalcsó Gyula</b>	
Képek és metaadataik gyűjteményezése scrapingtechnológiával közösségi képmegosztó oldalról .....	157
<b>Péter Róbert, Szántó Zsolt, Biacsi Zoltán, Kocsis Zoltán, Berend Gábor, Bilicki Vilmos</b>	
Az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz munkafolyamata és új funkciói .....	163
<b>Máray Tamás</b>	
Kvantum-számítástechnika: ez már a „jövő”?.....	171
<b>Fellegi Zsófia</b>	
Digitális kiadások migrációja: gépi és emberi intelligencia együttműködése.....	177
<b>Palkó Gábor</b>	
Posztmodern intertextualitás és digitális szövegkiadás .....	184
<b>Antal Dániel</b>	
A szlovák adatkicserélési tér magyarországi föderációjának lehetőségei.....	192
<b>Vass Johanna</b>	
Kutatási adatok megosztása a gyakorlatban – Adatrepozitóriumok használata az Ökológiai Kutatóközpont publikációiban .....	199
<b>Mihály Eszter, Micsik András, Nagy Kadosa</b>	
Irodalmi levélváltások nyomában TEI-vel és térképpel.....	208

## Az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz munkafolyamata és új funkciói

### The workflow and new functions of the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) multilingual research tool

Péter Róbert, Szántó Zsolt, Biacsi Zoltán, Kocsis Zoltán,  
Berend Gábor, Bilicki Vilmos  
*Szegedi Tudományegyetem*

[robert.peter@ieas-szeged.hu](mailto:robert.peter@ieas-szeged.hu)

#### Absztrakt

E rövid tanulmány célja, hogy bemutassa az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz működéséhez kapcsolódó munkafolyamatot és a különböző elemzőfunkciókat. A webes alkalmazás segítségével nagy mennyiségű metaadatot és szöveget lehet feldolgozni és kritikusán elemezni adatvezérelt mesterséges intelligenciával és természetesnyelv-feldolgozós technológiákkal támogatott módszerekkel és eszközökkel. Az egyszerűen használható felület interaktív paraméterbeállítást és -vezérlést biztosít a normalizálást is támogató előfeldolgozástól az analitikai szakaszokig. A felhasználók interaktív módon kísérletezhetnek az elemzések különböző beállításával a munkafolyamat során. A skálázható felhős infrastruktúrájának köszönhetően képes előfeldolgozni, elemezni és (szemantikailag) gazdagítani nagy mennyiségű szöveget és metaadatot számos nyelven. A beépített funkciók lehetőséget biztosítanak a szoros (*close*) és távoli (*distant*) olvasásra egyaránt. Az AVOBMAT felhasználóbarát, interaktív grafikus felületen integrál metaadat- és szövegelemzéssel kapcsolatos kutatási eszközöket.

**Kulcsszavak:** szöveg- és adatbányászat, természetesnyelv-feldolgozás, többnyelvű kutatási eszköz, metaadat, szemantikus adatgazdagítás

#### Abstract

The objective of this short paper is to introduce the workflow of the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) multilingual research tool, which enables researchers to critically analyse bibliographic data and texts at scale with the help of data-driven methods supported by Natural Language Processing (NLP) techniques. This exploratory tool offers a range of dynamic text and data mining tasks and provides interactive parameter tuning and control from the preprocessing to the analytical stages. It can preprocess, analyse and (semantically) enrich a vast number of texts and metadata in several languages due to its scalable infrastructure. The implemented analytical and visualization tools provide close and distant reading of texts and bibliographic data. The tools combine bibliographic data and NLP research methods in one integrated, interactive, user-friendly web application, allowing users to ask complex research questions.

**Keywords:** text and data mining, multilingual digital tool, natural language processing, metadata, semantic enrichment

E rövid tanulmány célja, hogy bemutassa az AVOBMAT többnyelvű szövegbányászati eszköz munkafolyamatát és új funkcióit. [7], [8], [9].<sup>1</sup>

## 1. A korpusz feltöltése

Az adatbázisokat feltölthetjük többféle formátumban. Az AVOBMAT automatikusan importálja a Zotero-gyűjteményeket CSV- és RDF-formátumban (teljes szöveggel), valamint a számos könyvtárban használt Eprints-es adatbázisokat (EP3 XML a teljes szövegek URL-jeivel). Az adatbázisokat feltölthetjük egy egyszerű CSV-fájl segítségével is. A teljes szövegek hozzáadása többféleképpen történhet: (i) a szövegek rögzíthetők egy külön erre a célra létesített CSV-s mezőben; a szövegekre mutató (ii) relatív vagy (iii) internetes útvonalat is megadhatjuk egy másik mezőben. A második opció esetében a teljes szövegeket tartalmazó mappát és a metaadatokat tartalmazó CSV-fájlt tömörítve kell feltöltenünk. TEI XML fájlok feldolgozásának fejlesztése folyamatban van. Az AVOBMAT minden olyan szövegformátumot tud importálni, amelyet az Apache Tika program képes kezelni, mivel ez alakítja át egyszerű szöveggé a bemeneti fájlokat.

## 2. A korpusz tisztítása

Az AVOBMAT számos lehetőséget kínál a szövegtisztításra. A helyettesítés (replace) funkció segítségével az optikai szövegfelismerésből (Optical Character Recognition: OCR) adódó hibákat javíthatunk, összevonhatunk szinonimákat, vagy modernizálhatjuk a régi, nem standardizált helyesírást használó szövegeinket. A cserepárok megadása során reguláris kifejezéseket is használhatunk. Így például lehetőségünk van különleges karakterek törlésére, rövidítések feloldására, az elválasztott szavak összevonására.

Ha nem szeretnénk a teljes szövegállományt elemezni, akkor a kontextusszűrő funkció segítségével megadhatunk kulcsszavakat, valamint a kulcsszavaktól balra és jobbra (külön-külön) található szavak számát. A későbbi elemzések során az AVOBMAT csak az így definiált szövegdobozokban található szavakat elemzi.

## 3. Paraméterek beállítása

Az előfeldolgozási fázisban a felhasználó konfigurálhatja az egyes elemzőket, valamint a metaadat- és szemantikus gazdagítást végző eszközöket. A munkafolyamat első lépése a feltöltendő szövegekre épülő automatikus nyelvdetekció, melynek eredményét az elemzések során figyelembe veszi a program.

---

<sup>1</sup> A tanulmány megjelenését a Szegedi Tudományegyetem Szegedi Tudományegyetem Interdiszciplináris Kutatásfejlesztési és Innovációs Kiválósági Központ Humán- és Társadalomtudományi Klasztere támogatta. Köszönetet mondunk a HUN-REN Cloud (lásd: Héder et al. 2022; <https://science-cloud.hu/>) használatáért, ami hozzájárult a publikált eredmények eléréséhez. A dolgozat tervezett publikálásának időpontjában az alkalmazás elérhető lesz a Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen szerverén a német Text+ humán- és társadalomtudományi infrastruktúra pilot projektje keretében. Lásd <https://avobmat-gwdg.web.app/app/instance/general/login> és <https://avobmat.hu/>

A felhasználók különböző konfigurációkat hozhatnak létre az egyes elemzésekhez, ahol az eredmény a szövegek nyelvétől függ. Kétféleképpen lehet egy dokumentumhoz nyelvet rendelni: a kutatók kiválaszthatják a feldolgozandó korpusz nyelvét (52 nyelv közül), vagy akár használhatják az automatikus nyelvfelismerés opciót is. Az utóbbi esetében a rendszer minden egyes dokumentumhoz önállóan választ nyelvet. A nyelv alapján a rendszer a spaCy-könyvtárra támaszkodó stopszó- és írásjel szűrést, valamint lemmatizálást is kínál.<sup>2</sup> Egyéni stopszó- és írásjel-listák is használhatók. A lemmatizáláshoz spaCy nyelvi modelleket használja a program, a spaCy által nem támogatott nyelvek esetében pedig LemmaGent [4].

A fentiek mellett az AVOBMAT-ban többek között a következő előfeldolgozási opciókat használhatjuk: a spaCy nyelvi modell kiválasztása (kis-, nagy- vagy transzformermodell); a szöveg kisbetűssé alakítása; a számok eltávolítása; minimális karakterhossz megadása.

A metaadatok bővítése magában foglalja a szerzők nemének azonosítását (férfi, nő, ismeretlen nemű vagy szerző nélküli) és a feltöltendő szövegekre épülő automatikus nyelvfelismerést. Az AVOBMAT-ba beépített női és férfi keresztnév-adatbázisok tartalmát a felhasználó bővítheti saját női és férfi névlistáival is, amelyek kiegészítik, és adott esetben felülírják a program által valószínűsített nemi kategóriákat.

A témamodellezés kapcsán a felhasználónak lehetősége van arra is, hogy a dokumentumokat azonos méretű szakaszokra darabolja. A felhasználók bizonyos lexikálisgazdagságelemzők (MSTTR, MATTR) esetében megadhatják a „szöveglablak” méretét (szószám) is paraméterként [2].

#### 4. A beállítások tesztelése és validálása

Az előző pontban említett konfigurációk megadása után az AVOBMAT megtisztítja és előfeldolgozza a feltöltött adatbázis egy részét. A felhasználó ily módon ellenőrizheti, hogy a beállított paraméterek megfelelőek-e mielőtt elindítaná a teljes adatbázis előfeldolgozását. Ha a konfiguráció megfelelőnek bizonyul, a beállítások elmenthetők egy sablonba. Ha a paraméterek további finomhangolására van szükség, a felhasználó akár újra is indíthatja a tisztítási és előfeldolgozási folyamatokat. Az AVOBMAT azonosítja a metaadatokban lévő hiányosságokat (érték nélküli mezők), ami az eredmények kritikus elemzését segíti.

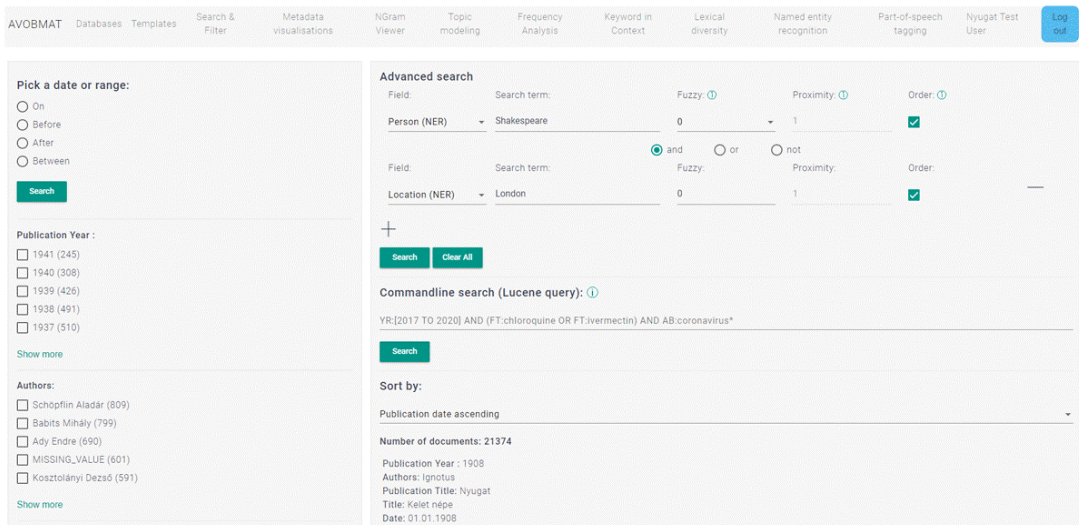
#### 5. Keresés és szűrés

A felhasználó kereshet és szűrhet a metaadatokban és a szövegekben fazettás, összetett és parancssori módokban. A szűrt adathalmazon végezheti el az összes elemzést. A dokumentumok NLP-és elemzésével gazdagítja a program a metaadatokat is. Például a felismert névelemek, mint például a személynevek, minden típusú keresésben és szűrésben megjelennek opcióként. Az AVOBMAT támogatja a paraméterezhető közelítő (*fuzzy*), szószomszédsági (*proximity*) kereséseket, valamint a parancssori lekérdezéseket is.

---

2 <https://spacy.io/usage/models>

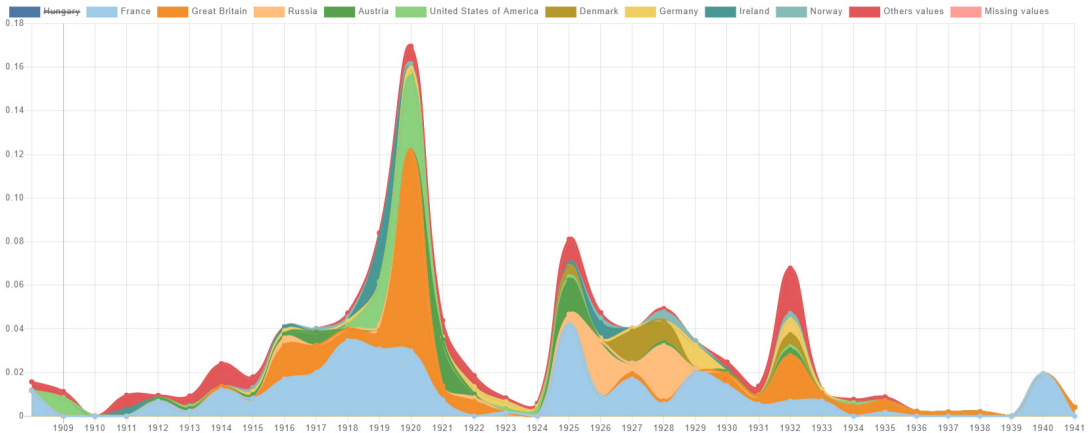




1. ábra: Az AVOBMAT grafikus felülete

## 6. Interaktív metaadat-elemzés

A felhasználók elemezhetik és vizualizálhatják a bibliográfiai adatokat (i) kronologikusan, vonal- és területdiagramokon, normalizált és aggregált formátumban; (ii) interaktív hálózati elemzést készíthetnek legfeljebb három metaadatmező segítségével; (iii) a megadott paraméterek alapján tetszés szerinti kör-, sáv- és oszlopdiagramokat készíthetnek a bibliográfiai adatok felhasználásával.



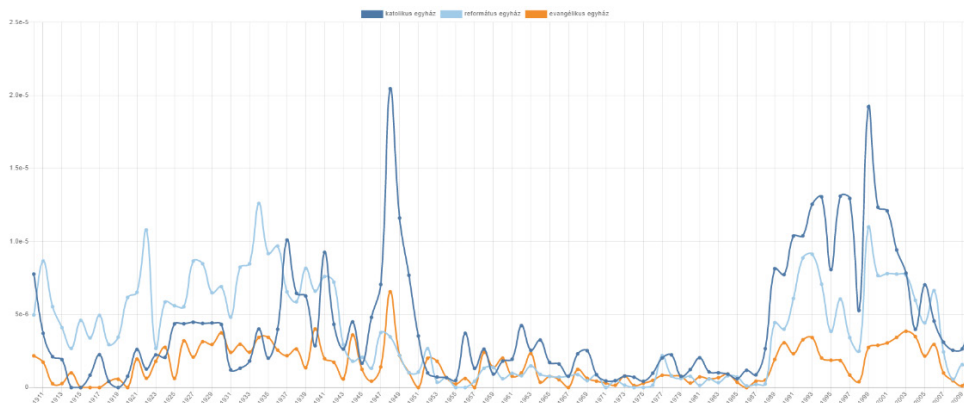
2. ábra: A Nyugat folyóirat külföldi szerzőinek eloszlása normalizált nézetben

## 7. Interaktív szövegelemzés

Az interaktív szövegelemzéshez a következő lehetőségek állnak rendelkezésre.

### 7.1. N-gram elemzés

A szövegek diakronikus elemzését az AVOBMAT n-gram elemzője támogatja. Idősoron megjeleníti a felhasználó által megadott – teljes szövegben található – n-gramok (itt egymás után következő n darab szó) éves eloszlását aggregált és normalizált módon.



3. ábra: A „katolikus egyház”, „református egyház” és „evangélikus egyház” bigramok eloszlása a Délmagyarország napilapban, 1911–2009

### 7.2. Szóstatistikai elemzések

A szófelhők hatékony eszközök lehetnek egy korpuszban valamilyen szempontból prominens szavak kiemelésére. Háromféle szóstatistikai elemzőt integráltunk az AVOBMAT alkalmazásba. A legegyszerűbb vizualizáció a szógyakoriság alapján készíti el a szófelhőt és mutatja az egyes szavakhoz tartozó gyakorisági adatokat. A második elemző (*Significant text*) azt mutatja, hogy milyen, az átlagostól jelentősen eltérő gyakoriságú szavak különböztetik meg egy digitális gyűjtemény általunk szűrővel kiválasztott részhalmozát a korpuszban található összes szövegtől [5], [10].<sup>3</sup> A harmadik elemző (*TagSpheres*) lehetővé teszi a felhasználók számára, hogy egy szó kontextusát vizsgálják [3]. A különböző szófelhők mellett a szóstatistikai adatokat oszlopdiagramokban is láthatjuk, és az itt szereplő adatsorokat exportálhatjuk.

### 7.3. Lexikai gazdagság

A lexikaidiverzitás-elemző segítségével nyolc metrika szerint tudjuk kiszámoltatni az egyes szövegek lexikai gazdagságát: Type-token ratio (TTR), Guiraud (Root TTR), Herdan (Log TTR), Mass TTR, Mean Segmental TTR (MSTTR), Moving Average TTR (MATTR), Measure of Textual Lexical Diversity (MTLD) és Hypergeometric Distribution Diversity (HDD) [2],[6],[11].

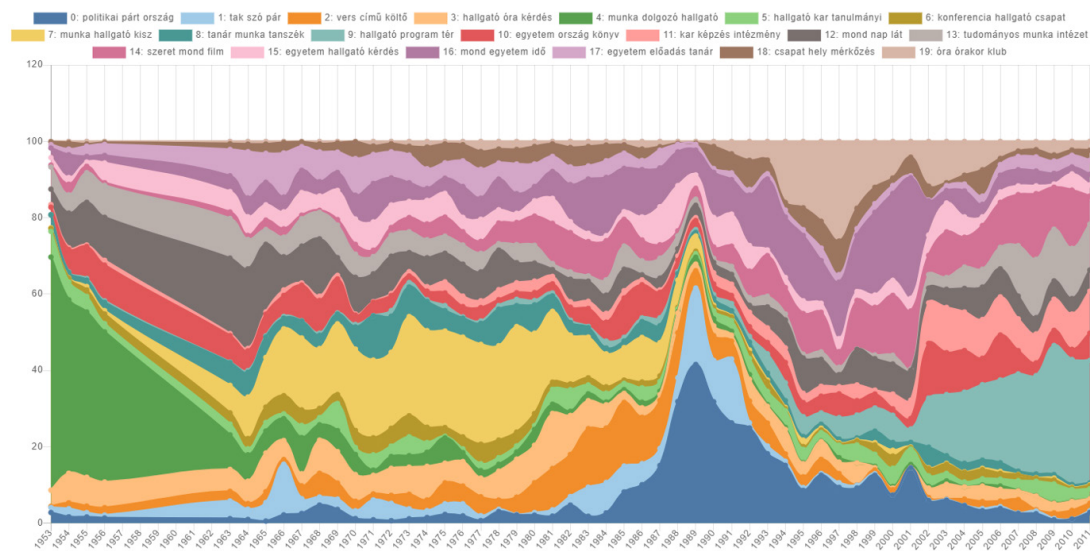
3 <https://www.elastic.co/guide/en/elasticsearch/reference/8.0/search-aggregations-bucket-significant-text-aggregation.html>

#### 7.4. Konkordancia

A konkordancia (*keyword-in-context*) eszköz segíti az elemezni kívánt szövegek szoros olvasását. Megadhatjuk, hogy az adott keresési kifejezés (akár több szó) környezetében hány betűt jelenítsen meg a program, valamint azt is, hogy maximum hány találatot mutasson.

#### 7.5. Témamodellzés

A témamodellzés segítségével rejtett és absztrakt témákat, szemantikai információkat fedezhetünk fel szövegekben. Az algoritmus statisztikai módszereket használ a szövegekbe ágyazott témák feltárására, valamint e témák kapcsolatainak és időbeli változásainak feltárására. Az AVOBMAT rendelkezik Latent Dirichlet Allocation (LDA) funkcióval [1], amely a felhasználó által megadott számú látens témát azonosít, ahol minden dokumentum e témák keverékének tekinthető. A módszer az együtt előforduló szavakat csoportosítja témákba, a dokumentumokhoz pedig valószínűségekkel hozzárendeli az egyes témákat. A modellezés eredményeit különböző módon tudja megjeleníteni az AVOBMAT. Megmutatja az egyes témákhoz kapcsolódó legrelevánsabb szavakat és dokumentumokat, megjeleníti e témák eloszlását idősoron, vizualizálja a különböző témák közötti korrelációkat. A felhasználók interaktív módon eltávolíthatják a stopszavakat.



4. ábra: A Szegedi Egyetem magazin témamodellzése

#### 7.6. Szófaji egyértelműsítés

Az AVOBMAT a spaCy nyelvi modellek segítségével 9 nyelven azonosítja a szófajokat. Az eredményekről ezután különböző interaktív vizualizációkat és statisztikai táblázatokat készít.

Word Form	Lemma	Part-of-speech Tag	Relative Frequency	Count ↓	Documents
életet	élet	NOUN	0.078810409	106	63
teszi	tesz	VERB	0.078810409	106	78
dolog	dolog	NOUN	0.078810409	106	73
Bang	Bang	PROPER NOUN	0.078810409	106	1
semmit	semmi	DETERMINER	0.078066914	105	63

5. ábra: Szófaji egyértelműsítés magyar nyelven

### 7.7. Névelem-felismerés, egyértelműsítés és összekapcsolás

A program jelenleg 16 nyelven azonosítja a névelemeket, például személyeket és helyneveket. A felismert névelemek száma és típusa nyelvenként eltérő. Az AVOBMAT különböző statisztikai táblázatokat és vizualizációt készít ezekről a névelemekről. Az utóbbiak a teljes szövegben is megtekinthetők. Jelenleg öt nyelv (angol, német, spanyol és portugál) esetében, a program egyértelműsíti a névelemeket és összekapcsolja ezeket a Wikidata-, VIAF- és ISNI tudásbázisokkal.

The screenshot shows a text document with several words highlighted in colored boxes, indicating their part-of-speech or category. A pop-up information box is visible for the name 'Robert Hanssen'. The box contains the following information:

- Robert Hanssen PERSON**
- "Robert Philip Hanssen" (born April 18, 1944) is an American former [[Federal Bureau of Investigation]] (FBI) [[double agent]] who spied for [[Soviet Union|Soviet]] and Russian intelligence services against the United States from 1979 to 2001. His [[espionage]] was described by the [[United States Department of Justice|Department of Justice]] as "possibly the worst intelligence disaster in U.S. history." Hanssen is currently serving 15 consecutive [[life imprisonment|life sentences]] without parole at [[ADX Florence]], a federal [[supermax prison]] near [[Florence, Colorado]].
- Wikidata
- ISNI
- VIAF

6. ábra: Névelem-felismerés és összekapcsolás az AVOBMAT-ban

## 8. Eredmények és konfigurációk exportálása, adatbázisok közzététele

Az AVOBMAT-tal végzett kísérletek és eredmények reprodukálhatóságát és átláthatóságát növeli a paraméterbeállítások JSON-formátumban történő importálásának és exportálásának lehetősége. A felhasználók a grafikus felületen sablonokat hozhatnak létre az előfeldolgozási és elemzési funkciókhoz. Az elvégzett elemzések statisztikai adatai és vizualizációi PNG- és különböző CSV-formátumokban menthetők, beleértve a Gephi-be importálható dokumentum-téma gráf fájlt a témamodellzés esetén. Ez utóbbi lehetővé teszi a kutatók számára, hogy a generált adatokat más szoftverekben használják fel. A felhasználók megoszthatják és nyilvánossá tehetik adatbázisukat.

## Irodalomjegyzék

- [1] Blei, D M, Ng, Y A. and Jordan, M I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003): 993–1022.
- [2] Covington, M A and McFall, J D. (2010) Cutting the Gordian Knot: the Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics* 17 (2): 94–100, <https://doi.org/10.1080/09296171003643098>
- [3] Jänicke, S and Scheuermann, G. (2017) On the Visualization of Hierarchical Relations and Tree Structures with TagSpheres. In: Braz, J et al. (eds.). *Computer Vision, Imaging and Computer Graphics Theory and Applications*. Cham: Springer International Publishing, pp. 199–219.
- [4] Juršič, M et al. (2010) Lemmagen: Multilingual Lemmatisation with Induced Ripple-down Rules. *Journal of Universal Computer Science* 16 (9): 1190–1214.
- [5] Manning, C D, Raghavan, P and Schütze, H. (2009) *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- [6] McCarthy, P M and Jarvis, S. (2010) MTL D, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behaviour Research Methods* 42 (2): 381–392, <https://doi.org/10.3758/BRM.42.2.381>
- [7] Péter, R, Szántó, Zs, Seres, J, Bilicki, V, and Berend, G. (2020) AVOBMAT: a digital toolkit for analysing and visualizing bibliographic metadata and texts. In: Berend, G, Gosztolya, G and Vincze, V (eds.) XVI. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem, Informatikai Intézet, pp. 43–55.
- [8] Péter, R, Szántó, Zs, Seres, J, Bilicki, V, and Berend, G. (2022) Az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz bemutatása. *Digitális Bölcsészet* 4 (2022): 3–28, <https://doi.org/10.31400/dh-hun.2021.4.3530>
- [9] Péter, R.; Szántó, Zs; Biacsi, Z.; Berend, G.; Bilicki, V. (2024). “Multilingual Analysis and Visualization of Bibliographic Metadata and Texts with the AVOBMAT Research Tool.” *Journal of Open Humanities Data* 10: 1–10.
- [10] Rudi, L C and Vitányi, P M B. (2007) The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering* 19 (3): 370–383, <https://arxiv.org/pdf/cs/0412098v3.pdf>
- [11] Torruella, J and Capsada, R. (2013) Lexical Statistics and Tipological Structures: A Measure of Lexical Richness. *Procedia: Social and Behavioral Sciences* 95 (2013): 447–454, <https://doi.org/10.1016/j.sbspro.2013.10.668>