Doménspecifikus korpusz építése és validálása

Dodé Réka

ELTE BTK Nyelvtudomány Doktori Iskola
Alkalmazott nyelvészet program

2017. február 3.

Áttekintés

1. Bevezetés és hipotézisek
2. Filtereket alkalmazó keresőrobot
3. A validálás módszere
4. Eredmények és következtetések

Bevezetés és hipotézisek

- Terminuskivonatoló alkalmazás
- Doménspecifikus korpusz
- Ökoinnováció doménhez tartozó szövegek
- Hasonlóság: ugyanahhoz a doménhez tartozó szövegek

Bevezetés és hipotézisek

1. Hipotézis

A letöltött szövegek elérik a 0,50-es hasonlósági értéket.

2. Hipotézis

Minél kevesebb szövegszóból áll a szöveg, annál kevesebb lesz a hasonlósági értéke a referenciaszöveghez képest.

Filtereket alkalmazó keresőrobot

- Filtereket alkalmazó keresőrobot (Grefenstette–Muchemi 2016)
- Kezdő URL-lista (40 db)
- Mintákat tartalmazó fájl (kézzel: 48 kifejezés)
- Minta alapján válogatott szövegek
- 2.000 fájl, 2.625.164 szövegszó

Validálás módszere

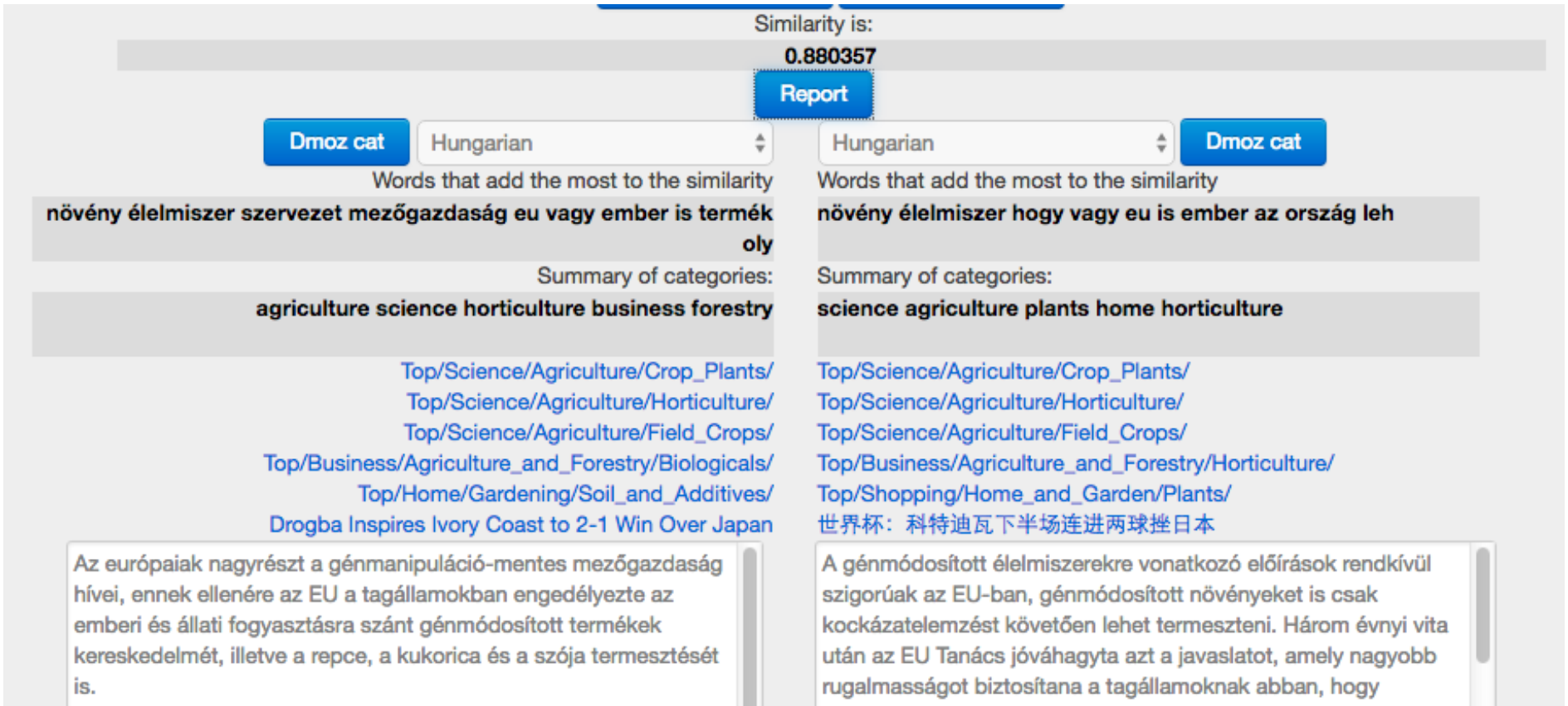
- 20 db véletlenszerűen kiválasztott szöveg (min. 500 szövegszó).
- Referenciaszövegek kiválasztása – 4 db.
- Választás minták alapján.
- Lemmatizálás és stopszavak kiszűrése.
- Hasonlóság mérése: XLike projekt kereteiben készült összehasonlító program.

JSI Similarity Service

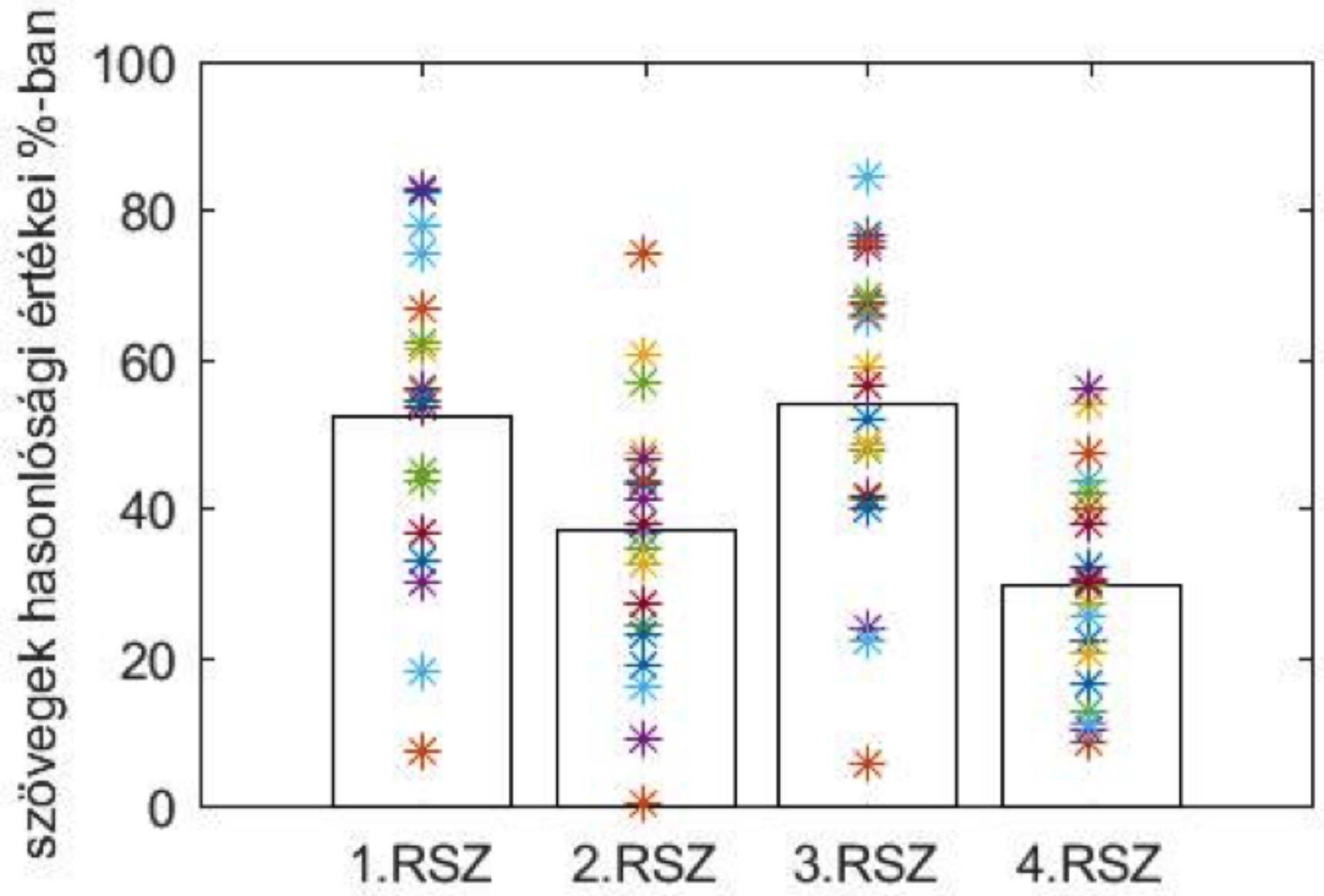
- Dokumentumindexelés fogalomalapú megközelítéssel.
- Vektortér-modellezés: vektor reprezentálja a dokumentumokat a vektortérben.
- Indexelés rejtett szemantikai elemzéssel: a megfigyelt vektorokon kívül rejtett tulajdonságok is léteznek.
- Cél: kiválasztani a szöveg néhány olyan jellemző attribútumát, amely még alkalmas a dokumentumok közelítő jellemzésére és páronkénti összehasonlítására.
- Előny: a szinonimák kezelése (a dimenzióredukcióval az azonos jelentésű szavak hasonló helyen jelennek meg a térben).

JSI Similarity Service

- Hasonlóság számolása a szövegek között: vektorok hasonlóságának mérése.
- Koszinusz hasonlóság mérése.
- A hasonlóság mérésekor kapott értékek (-1) 0 és 1 közé esnek.



Eredmények I.



Eredmények II.

A 20 random kiválasztott szöveg átlag hossza stopszavak nélkül **1428 szövegszó**, a szövegek hosszának szórása **1054,49**.

Referenciaszöveg	1	2	3	4
Átlag hasonlóság	0,524	0,373	0,541	0,297
Szórás	0,198	0,192	0,201	0,14

Table 1: Átlag és szórás referenciaszövegenként.

Az értékek nem normál eloszlásúak, ezért non-parametrikus tesztet (az egymintás t-próba nem parametrikus megfelelőjét) használtunk annak megállapítására, hogy a hasonlósági értékek eltérnek-e az előre megállapított 0,50-es küszöbtől.
Egyik szöveg esetében sem találtunk szignifikáns eltérést az 0,50-es értéktől ($p_{1\text{szöveg}}$ =0,68; $p_{2\text{szöveg}}$=0,36; $p_{3\text{szöveg}}$=0,80; $p_{4\text{szöveg}}$=0,22).

Eredmények III.

Referenciaszövegek	1–2	1–3	1–4	2–3	2–4	3–4
Hasonlóság	0,585	0,844	0,443	0,772	0,568	0,505

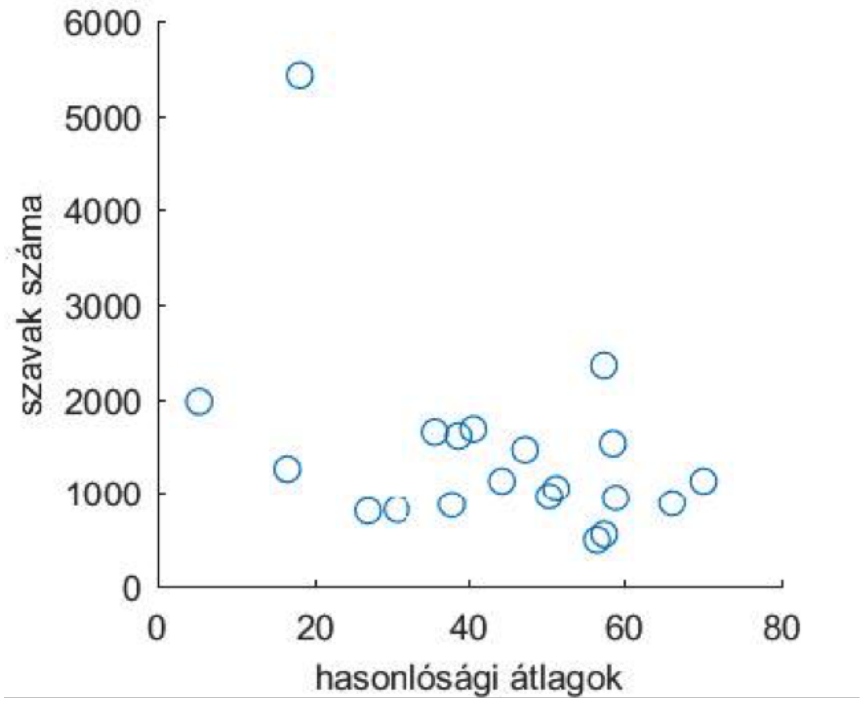
Table 2: Referenciaszövegek hasonlósága egymáshoz mérve.

Referenciaszövegek		p-értékek	szignifinkáns (<0.05*, <0.01**, <0.001***)
1.	2.	0,021	*
1.	3.	0,695	
1.	4.	<0,001 (0,0006)	***
2.	3.	0,011	*
2.	4.	0,223	
3.	4.	<0,001 (0,0004)	***

Table 3: Szövegek hasonlóságának eltérése referenciaszövegenként.

Páros t-próba nem parametrikus formája

Eredmények IV.



A szövegek szószámának kapcsolata a hasonlóság mértékével (korrelációs analízis).
Enyhe tendenciózus ($0.05<p>0.1$) negatív kapcsolat van közöttük ($r_{Pearson}$=-0.41, p=0.071) → egyetlen kiugró szószámmal rendelkező szöveg okozta ($r_{Pearson}$=-0.25, p=0.30).

Következtetések I.

1. Hipotézis

Az első hipotézis részben teljesült, mivel egyik referenciaszöveghez mért átlag hasonlósági érték sem lett magasabb, mint 0,50.

2. Hipotézis

A második hipotézis nem teljesült, mivel a szövegek mérete és az eredmény között bár van kapcsolat, de egy gyenge negatív kapcsolat, amit egy szöveg okozott.

Következtetések II.

1. A szkript egyszerűen átültethető más nyelvekre, robosztus.
2. Előzetesen nincs szükség a szövegek elemzésére.
3. A minták kiválasztása kulcsfontosságú.
4. Validálásra így is szükség van.
5. A további statisztikai vizsgálatok a minta növelését igénylik.

További vizsgálatok – Kitekintés

- A korpusz összeállításánál – minták összeállításánál – figyelembe kell venni a ökoinnováció aldoménjeit (pl. megújuló energia, újrahasznosítás).
- Emberi annotátorok bevonása: a szöveg milyen mértékben tartozik a doménhez, illetve miért.

Irodalom

- Rettinger, A. – Zhang, L. – Rupnik, J. – Muhič, A. 2012. *Cross-lingual document linking prototype.* Deliverable D4.1.1. XLike project.
- Grefenstette, G. – Muchemi, L. 2016. *Determining the Characteristic Vocabulary for a Specialized Dictionary using Word2vec and a Directed Crawler.* Lexicographic Resources for Human Language Technology GLOBALEX 2016 Workshop Proceedings, 24 May 2016, Portoroz.
- Nagy Á. 2012. *Terminológiakivonatolás francia nyelvű szabadalmi leírásokból szabály alapú és statisztikai módszerek segítségével.* PhD-értekezés. Szegedi Tudományegyetem Bölcsészettudományi Kar, Nyelvtudományi Doktori Iskola.
- Remus, S. – Biemann, Ch. 2016. *Domain-Specific Corpus Expansion with Focused Webcrawling.* Proceedings of the 10th edition of the Lexicographic Resources for Human Language Technology 23-28 May 2016, Portorož (Slovenia).
- García–Cuesta, E. – Galán, F. – Muhic, A. – Trampus, M. – Li, Zh. – Carreras, X. 2013. *Early Prototype.* D6.2.1. XLike project.
- Lu, J. – Ruan, D. – Zhang, G. 2007. *E-Service Intelligence. Methodologies, Technologies and Applications.* Springer-Verlag, Berlin Heidelberg.

Köszönöm szépen a figyelmet!

kovacs.reka@nytud.mta.hu