

Mi lehet a válasz?

Különböző hosszúságú szövegek tartalmának az összehasonlítása



Pázmándi Eszter
ConTEXT, 2019. november 26.

PROBLÉMA

- Kereskedelmi bank helpdesk-je
 - Leterheltség
 - Nagy és kezelhetetlen mennyiségű szöveges dokumentum/adat
 - Feleslegesen feladott hibajegyek növekvő száma
 - Ismétlődő hibák

MEGOLDÁS?

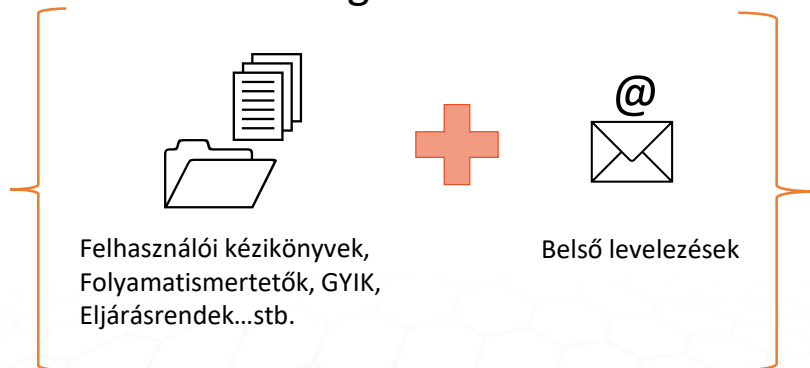
- Tudásbázis létrehozása
 - Dokumentumok közötti hatékony keresés
 - Információk gyors megtalálhatósága
 - Információk szelektív közvetítése
 - Szabály-alapú rendszer

KIHÍVÁS

- Kérdés: szabadszavas és rövid
- Válasz: kötött és hosszú
- Kérdés és válasz tartalmának az összehasonlítása
 - Különböző hosszúság
 - Kérdésben kevesebb szó → releváns válasz (itt: dokumentum-részlet/fejezet) megtalálása

FOLYAMAT ÁTTEKINTÉSE

Szöveges források



Szövegek feldolgozása

- Strukturális felbontás
- Tisztítás
- Témák szerinti csoportosítás szavak súlyozott gyakorisága alapján (TF-IDF)



Szótárépítés

- Típusok létrehozása
- „Hivatalos” nyelvezetre (dokumentumok)
- „Gyakorlati” nyelvezet (belső levelezések) → szinonima szótár



Szövegbányászat

- Kulcsszavak, kulcskifejezések kinyerése
- Kapcsolatok, mintázatok feltárása



Szövegek tartalmának az összehasonlítása

- Cosinus hasonlóság: szótár → kulcsszó-dokumentum mátrix alapján

SZÖVEGEK FELDOLGOZÁSA

- Szöveges források
 - Dokumentumokban elérhető információk (kb. 150 file)
 - Eljárásrend, Folyamatismertető, Felhasználói kézikönyv, Függelékek, mellékletek, GYIK, Oktatási anyagok
 - 150 fájl → 528 egység/fejezet
 - Email-ek
 - Témánként csoportosítva
 - Kérdés-válasz párok azonosítása



Corpus



Felhasználói kézikönyvek,
Folyamatismertetők, GYIK,
Eljárásrendek...stb.



Belső levelezések

SZÖVEGEK FELDOLGOZÁSA

- Típusok meghatározása: „alternatív” klaszterezés

- Egymást átfedő csoportok
- Nehezen elkülöníthető témák

➔ Szó-dokumentum mátrix + Belső levelezések téma szerinti bontásának a használata

- releváns szavak kinyerése fájlonként → fókusz: kulcsszavakon → emailek-ben megjelenő kulcsszavakkal való összehasonlítás → file-ok témába sorolása

Szavak gyakorisága



Szavak súlyozott gyakorisága



Szavak gyakorisága



Szavak súlyozott gyakorisága

SZÓTÁR ÉPÍTÉSE

- Releváns kategóriák/szavak/témák meghatározása
 - Corpusban előforduló szavak összegyűjtése kategóriák/témák szerint
 - pl.: szentiment szótár: érzelmek-/véleményhez köthető szavak gyűjtése → adott szöveg milyen érzelmet/véleményt fejez ki

pozitív	negatív
szép	ronda
finom	ízetlen
étvágygerjesztő	gusztustalan

➤ Szótárban szereplő szavakat fontosságuk alapján súlyozva tudjuk figyelembe venni

SZÓTÁR ÉPÍTÉSE

„Hivatalos” nyelvezet

- Dokumentumok, szabálykönyvek...stb.



„Gyakorlati” nyelvezet

- emailek

➔ **Szinonima szótár:**
- Adott szavak/kifejezések ragozott alakjai + szinonimái → nyelvezet, kérdésben használt szavak beazonosítása

Szótár mérete:

- 56 típus
- 572 szó/kifejezés
- 4373 szinonima

Típus = egy adott témához/tevékenységhez/folyam athoz kapcsolódó szavak csoportja



Kulcstípus

Szavak, kifejezések listája egy típuson belül

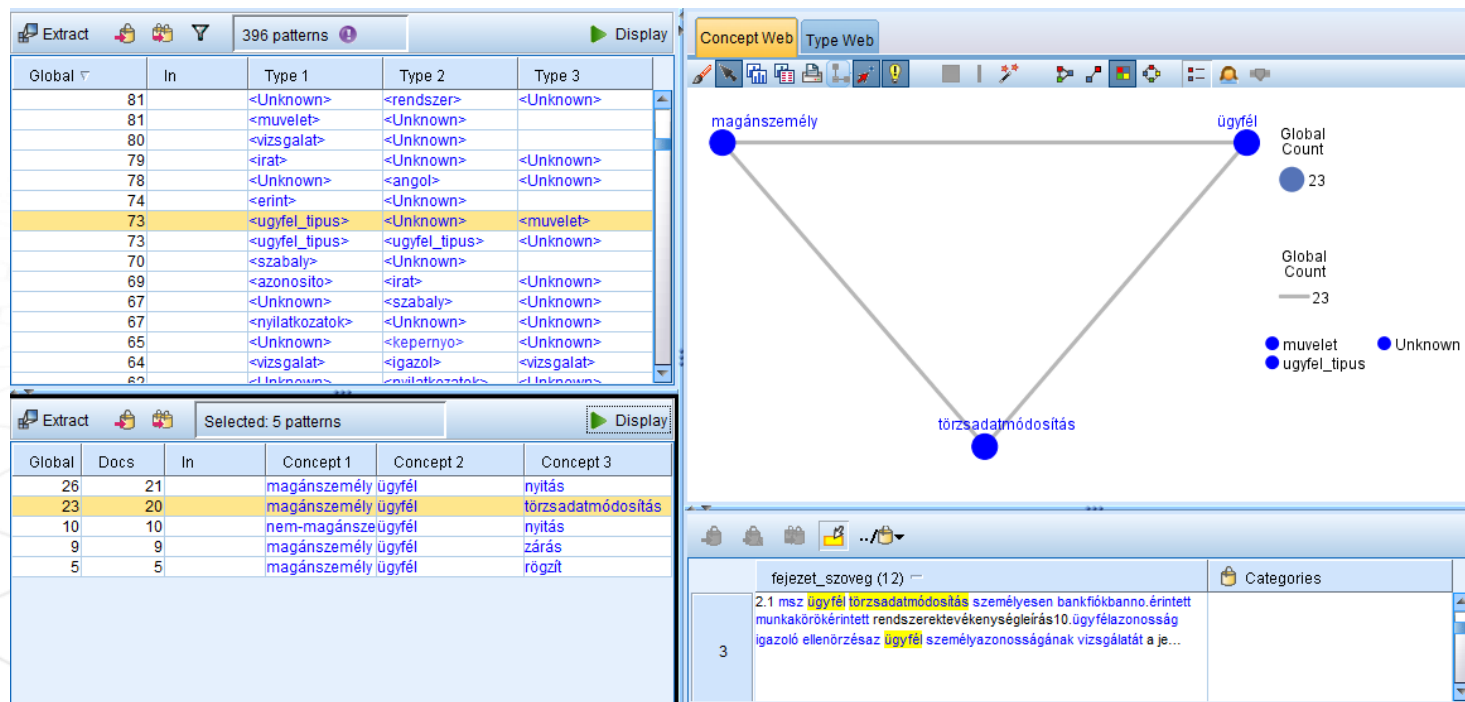
Term	Match	Inflect	Type	Library
gomb	Entire Term	☒	kepernyo	Local Library
kezdolap	Entire Term	☒	kepernyo	Local Library
klikkel	Entire Term	☒	kepernyo	Local Library
képernyő	Entire Term	☒	kepernyo	Local Library
képernyői	Entire Term	☒	kepernyo	Local Library
képernyőkép	Entire Term	☒	kepernyo	Local Library
legördülő	Entire Term	☒	kepernyo	Local Library
lenyíló	Entire Term	☒	kepernyo	Local Library
munkakosár	Entire Term	☒	kepernyo	Local Library
screen	Entire Term	☒	kepernyo	Local Library
screenshot	Entire Term	☒	kepernyo	Local Library
sáv	Entire Term	☒	kepernyo	Local Library

Target	Synonyms	Library
adatkészlet	adatkészletek	Local Library
adatkör	adatkörhöz, adatkört	Local Library
adatközlés	adatközlésre	Local Library
adatküldés	adatküldést	Local Library

1 Library 56 Types 572 Terms 70 Excludes 4373 Synonyms 0 Optional

SZÖVEGELEMZÉS

- Kapcsolatok, mintázatok feltárása
- Nem csak szavak
 - szabályok alkotása: 2-3 tagú kulcskifejezések, amelyek legalább egyik tagja szótári elem



id	Matched Text	kulcsszavak	kulcstipusok
205	ha a módosított adat nem érinti a <fatca> <*besorolást>, nem kell új nyilatkozat, d...	fatca besorolás	nyilatkozatok muvelet
205	ha a módosított adat nem érinti a fatca besorolást, nem kell új <nyilatkozat>...	nyilatkozat	irat_tipus
205	ha a módosított adat nem érinti a fatca besorolást, nem kell új nyilatkozat, de...		rendszer
205	ha a módosított adat nem érinti a fatca besorolást, nem kell új nyilatkozat, de...	kitöltés fatca mezőket	muvelet nyilatkozatok
205	ha a módosított adat nem érinti a fatca besorolást, nem kell új nyilatkozat, d...	korán kitöltött nyilatkozat	irat_tipus
205	ha a módosított adat nem érinti a fatca besorolást, nem kell új nyilatkozat, d...	adat alap	adat

HOGYAN HASONLÍTSUNK ÖSSZE SZÖVEGEKET?

*„A macska megette az ebédem.” / „A macska megette az ebédét.”
„A kutya a réten napozik.” / „Az eb a mezőn sütteti a hasát.”*

- Mennyire hasonló két szöveg
 - lexikai hasonlóság: a „felszínen”, szavak szintjén (egyes szavak előfordulása)
 - szemantikai hasonlóság: jelentés szintjén
- Különböző módszerek kombinálása:
 - Szótár
 - (Kulcs)Szó-dokumentum mátrix → kulcsszavak
 - Szavak gyakoriság (tf)
 - Inverz dokument gyakoriság (idf)
 - Cosinus hasonlóság: vektorizálás → két vektor cosinusa

MŰKÖDÉSI LOGIKA

Input



Kérdés



„Ha visszavonom a műveleti listából a feladatot, akkor azt a tableten is meg kell tennem?”

Szó-dokumentum mátrix

- Kulcstípusok & kulcsszavak
- szótár alapján

Corpus



Szó-dokumentum mátrix

- Kulcstípusok & kulcsszavak
- szótár alapján



Merge
&
Match



Cosinus
hasonlóság



Kulcsszavak
távolsága



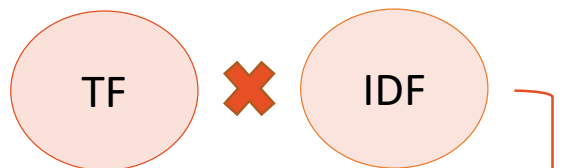
80% -os pontosság

(KULCS)SZÓ-DOKUMENTUM MÁTRIX

- Szótár alapján kinyert kulcsszavak vizsgálata
- Dokumentumok/szavak súlyozása
- TF-IDF: „Szöveg számokká konvertálása”
 - Kulcsszavak kinyerése, gyakoriságukkal normalizálva → szavak fontossága az elemzett szövegrészletben
 - Nem lehet logikai kapcsolatot felírni (→ távolság mérése)
 - TF: term frequency
Adott szó gyakorisága a dokumentumban / $\sum$ Szavak száma a dokumentumban
 - IDF (inverse document frequency): ritka kifejezés → magas idf ↔ gyakori kifejezés → alacsony idf
 $1 + \log_e(\sum \text{Dokumentumok száma} / \sum \text{Dokumentumok száma, amelyek tartalmazzák az adott szót})$

KULCSSZÓ-DOKUMENTUM MÁTRIX

Dokumentum



Kérdés



- Kérdés: **Gyűrűk ura**
- D1: „Egy gyűrű mind felett.”
- D2: „A Gyűrűk ura írója, Tolkien. ”
- D3: „A második könyv címe a Gyűrű szövetsége. ”

TF-IDF						
szavak	IDF dokumentumok (csak a keresett szavakra)	IDF kérdés	TF*IDF D1	TF*IDF D2	TF*IDF D3	TF*IDF kérdés
gyűrű	1,18	0,69	0,29	0,24	0,17	0,59
úr	1,18	0,69	0	0,24	0	0,59

Term frequency				
szavak	D1	D2	D3	kérdés
tolkien	0,25	0,2		
mind	0,25			
felett	0,25			
gyűrű	0,25	0,2	0,14	0,5
úr		0,2	0,14	0,5
író		0,2		
második			0,14	
könyv			0,14	
cím			0,14	
szövetség			0,14	

COSINUS HASONLÓSÁG

Kapott találatok és kérdés tartalmának az összehasonlítása: **Cosinus hasonlóság**

- Két szó/mondat/szövegrészlet hasonlóságának a mérésére használt mutatószám
- Cél: releváns dokumentumok megtalálása a feltett kérdésre
- Dokumentumok rangsorolása
- Két vektor közötti szög cosinus értéke
 - nagy szög → nagy távolság → cosinus → (-1)
 - (kis szög → kis távolság → cosinus → 1)

- **Kérdés: Gyűrűk ura**
- D1: „Egy gyűrű mind felett.”
- D2: „A Gyűrűk ura írója, Tolkien.”
- D3: „A második könyv címe a Gyűrű szövetsége.”

"Dot" termék * (Dokumentum, Kérdés)

$$\frac{|\sqrt{(\sum TF * IDF_{Dokumentum})^2} * |\sqrt{(\sum TF * IDF_{Kérdés})^2}|}{}$$



Cos(Dokumentum, Kérdés)

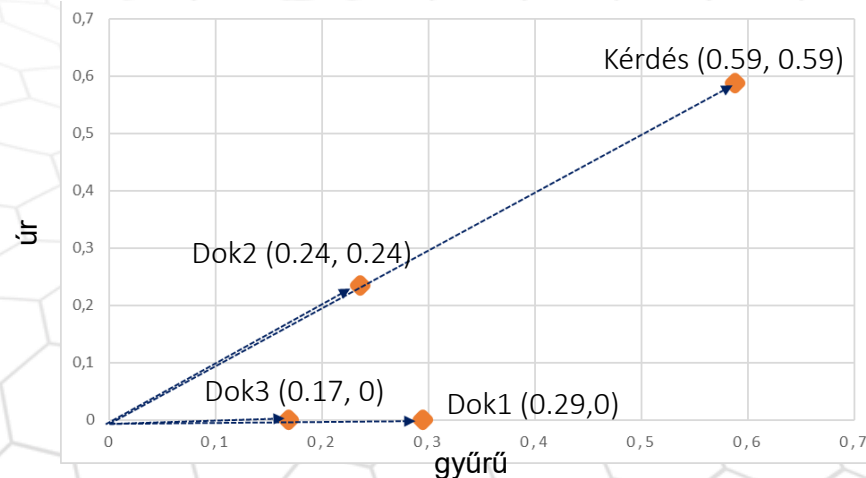
mondatok	"Dot"	Dok.	Kérdés
D1	0,17	0,29	0,83
D2	0,28	0,33	0,83
D3	0,10	0,17	0,83



Cos(Dokumentum, Kérdés)

D1	0,99
D2	0,77
D3	0,88

TF*IDF ábrázolása



»»» Kérdés: „Ha visszavonom a műveleti listából a feladatot, akkor azt a tableten is meg kell tennem?”



össz kulcsszó Max	Kulcsszavak	tf_query	Idf query	Tf-idf query	kulcstípusok	szavak össz (dok)	tf dok	mennyiben	idf dok	Tf- idf dok	kulcszavak száma	fejezet címe	match	dot	query	document
7	feladat	0.067	3.303	0.220	feladat	368	0.117	119	2.492	0.291	829	3.5. az ellenőrzési feladat	"... kontroll feladat is létrejön..."	0.064	0.048	0.085
7	tablet	0.067	3.303	0.154	egyeb_szavak	218	0.009	13	4.706	0.043	443	négy szem elves ellenőrzés	"...valamint az aláírásminta megadását az ügyfél adatlapon, vagy a tableten..."	0.007	0.024	0.002
7	visszavon	0.067	3.303	0.220	ugyfel_tevek enyseg	257	0.004	9	5.074	0.020	441	ügyfél bizalmi vagyonkezelő vállalkozás	"...az mnb az engedélyt visszavonja, ha..."	0.004	0.048	0.000



Fájl	Dot	Cosinus hasonlóság
F1	0.156	0.984
F2	0.012	0.680
F3	0.012	0.680

„...műveletek visszavonására csak biometrikus hitelesítés alatti státuszban van lehetőség a művelet végén található piros visszavonás ikonra kattintva. művelet/összevont műveletek visszavonása esetén a státusz újra rögzített lesz. a műveleti listából történő visszavonás nem jelenti a tableten lévő feladat automatikus elvetését is, azt külön a tablet felületén meg kell tenni!...”

ÖSSZEGZÉS

- Jelentős humánkapacitás-, idő- és tapasztalatigényes
- Hasonlóság alapú keresés
- Bármikor reprodukálható
- Konzisztens
- Más területen is alkalmazható
- Tesztelés alatt

Köszönöm a
figyelmet!