

Návod k programu

1 Načíst soubory Vektorová reprezentace LSA PLSA LDA Nezaokrouhlovat

2 TF TF - IDF TF - IDF+1 - Téma: 10 +

Březinka.csv	ELO.csv	HK.csv	Královehradec...	UHK.csv	fotbal.csv	kunětickáHora...	lkvz26.csv	obrana.csv	opevneni.csv	rock.csv	tvrze.csv
45.78983	17.58649	23.85541	46.75103	31.07347	21.42723	24.71453	17.99918	11.93553	25.22061	48.51985	27.42088
-21.85291	4.50866	1.5031	-4.11545	6.26978	-3.20715	-1.8666	-0.21817	-2.96792	-6.23466	35.10065	-18.54655
19.58637	3.20227	-8.13658	-28.35141	-12.70424	-3.48491	-1.94741	-0.07754	0.54874	2.88784	15.81309	8.70565
-8.00529	-5.27028	-2.45074	12.44654	-27.84286	9.57798	-1.80085	1.02281	2.14703	2.07168	7.26027	6.99179
11.82088	4.21564	-2.27916	13.29632	-8.47223	-22.05685	0.58307	-5.83184	-4.12959	-5.58082	1.76984	-9.19211
-9.3513	-8.39659	4.25528	-0.52725	3.72445	-16.10996	1.967	-3.47385	0.92803	1.8772	3.38487	18.95338
6.31783	-10.82723	-0.79619	3.831	5.42769	4.64815	-18.41391	-7.596	-2.50504	-6.1562	4.00166	2.02662
-3.9924	17.50291	-1.89913	1.87266	0.40977	-2.09377	-13.0013	3.54634	0.81025	0.56164	-1.70535	6.61105
1.30615	-7.10522	5.89291	-0.85927	-1.56268	-5.84151	-7.97913	15.81046	3.56301	7.46592	-0.53944	-5.59996
-0.60205	-3.27719	-17.09019	3.21641	4.08987	-0.84005	2.08992	8.71632	-2.72326	-2.15518	1.0613	2.19778

3

4

Podobnost dokumentů Březinka.csv a obrana.csv je: 0.65559 (pro novou klikněte)

1. Po spuštění programu bude viditelné pouze tlačítko *Načíst soubory* a *Nezaokrouhlovat*.

- Tlačítko *Načíst soubory* otevře nové okno s průzkumníkem souborů, ve kterém očekává vybrání vstupních souborů ve formátu CSV, souborů je nutné vybrat více najednou. Pokud načtení z nějakého důvodu selže, soubory mohou být načteny znovu a zobrazí se upozornění místo tabulky 3.
- Po načtení souborů se zobrazí *Vektorová reprezentace* kolekce dokumentů v tabulce 3 a zobrazí příslušná tlačítka v řadě 2.
- LSA* provede výpočet latentní sémantické analýzy a zobrazí výsledky v tabulce 3 a zobrazí příslušná tlačítka v řadě 2.
- PLSA* vypočte pravděpodobnostní latentní sémantickou analýzu a zobrazí výsledky v tabulce 3 a zobrazí příslušná tlačítka v řadě 2.
- LDA* provede výpočet latentní Dirichletovy alokace a zobrazí výsledky v tabulce 3 a zobrazí příslušná tlačítka.
- Nezaokrouhlovat* / *Zaokrouhlovat* přepíná zaokrouhlování výsledků na pět desetinných míst.

2. Tyto tlačítka budou viditelná pouze po načtení souborů a v závislosti na vybrané metodě.
 - a. *TF* přepne výpočet na četnost slov.
 - b. Podobně jako *TF*, ale matice bude vypočítána pomocí *TF-IDF*, tedy četnosti slov dělené inverzní četností dokumentů.
 - c. Viz předchozí případ, při výpočtu *TF-IDF* přičte k logaritmu inverzní četnosti dokumentů jedničku.
 - d. Tlačítka + a - zvýší nebo sníží počet témat pro metody *LSA* apod.
 - e. Tlačítko *Témat: X* provede znovu výpočet zvolené metody s aktuálně nastaveným počtem témat.
3. Tabulka zobrazující matici hodnot četnosti slov nebo výsledků některé z metod sémantické analýzy textů. Pokud nejsou načteny soubory, nebo probíhá výpočet, tabulka je zobrazena prázdná nebo pouze s hlavičkou a je zde umístěn informační text. Výsledky jsou zaokrouhleny podle stavu tlačítka.
4. *Podobnost dokumentů*. Tato položka se zobrazí po načtení dokumentů a pro její výpočet je potřeba dvakrát kliknout levým tlačítkem myši do tabulky 3 na dva sloupce zobrazující dokumenty. Podobnost je vypočítána pomocí kosinové podobnosti.

Program není nijak výpočetně optimalizován a algoritmy nejsou nijak zlepšeny pro větší kolekce dokumentů, takže s větší kolekcí dokumentů nebo s velkým počtem slov může výpočet trvat delší dobu a dá se zastavit pouze vypnutím programu.

Vlastní vstupní dokumenty musejí být ve formátu CSV s e středníkem jako oddělovačem a s daty ve druhém sloupci. Data nemusejí býti lemmatizovaná, ani česky psaná, ale výsledky jiných dat nemusejí mítí vypovídací hodnotu.