

labeldate1

slovakandothers = et al., and = a, in =

checklistitemize1 [checklist] label= $\square$, leftmargin=*, nosep

ilinská univerzita v iline

Fakulta riadenia a informatiky

Úvod do strojového uenia

Semestrálna práca . 2

Document topic discovery

Obsah

1	Cie	1
2	Matematická reprezentácia textu	1
3	Optimalizácia a redukcia umu	1
4	Zhlukovanie	1
5	Stanovenie granularity tém	2
6	Interpretácia objavených tém	2
7	Záver	3

Zoznam obrázkov

1	Závislos sumy tvorcov (WSS) od potu zhlukov k . Lake indikuje optimálnu granularitu tém.	2
---	--	---

1 Cie

Hlavným cieom je zisti, aké tematické oblasti sa prirodzene vyskytujú v kolekcii textov bez toho, aby boli vopred určené triedy alebo kategórie. Ide teda o úlohu uenia bez uitea. Na základe tatistického rozdelenia výrazov a podobností medzi dokumentmi zoskupíme lánky do niekorych zhluok, priom každý z nich by mal zodpoveda uritej téme.

2 Matematická reprezentácia textu

Na to, aby bolo moné texty matematicky spracova, je potrebné ich previes do vektorovej podoby. Vyuitá bola reprezentácia TF-IDF (*Term FrequencyInverse Document Frequency*), ktorá prirauje vyiu váhu slovám, ktoré sú pre daný dokument typické, ale nie príli asté.

Frekvencia slov (TF) sa poíta pomocou logaritmickej mierky:

$$\text{TF}(w, d) = \log(1 + f(w, d)) \quad (1)$$

kde $f(w, d)$ oznauje poet výskytov slova w v dokumente d . Inverzná frekvencia dokumentov (IDF) vyjadruje, ako informane je slovo prínosné vzhľadom na svoj výskyt v súbore dokumentov D :

$$\text{IDF}(w, D) = \log\left(\frac{N}{f(w, D) + 1}\right) \quad (2)$$

kde N predstavuje celkový poet dokumentov. Finálna hodnota TF-IDF je daná súinom týchto dvoch lenov.

3 Optimalizácia a redukcia umu

Vektory TF-IDF obsahujú aj znany um a redundantné informácie, o môe negatívne ovplyvni kvalitu zhlukovania. Preto bola pouitá metóda hlavných komponentov (PCA), aby sa znížili rozmery dát a zároveň zachovali o najviac variabilitu.

PCA premieta pôvodné dáta do priestoru hlavných komponentov, ktoré sú lineárne nezávislé a postupne zachytávajú najväi rozptyl v dátach. Projekciou na prvých niekoľko komponentov sa odstránia menej informatívne rozmery a tým sa model zjednoduú.

4 Zhlukovanie

Na samotné zhlukovanie bol pouitý algoritmus K-means. Ide o iteratívny algoritmus, ktorý rozdeje priestor príznakov do k skupín a minimalizuje nasledujúcu funkciu nezhody:

$$L(\Theta) = \frac{1}{n} \sum_{i=1}^n \min_{m=1, \dots, k} \|x_i - \theta_m\|_2^2 \quad (3)$$

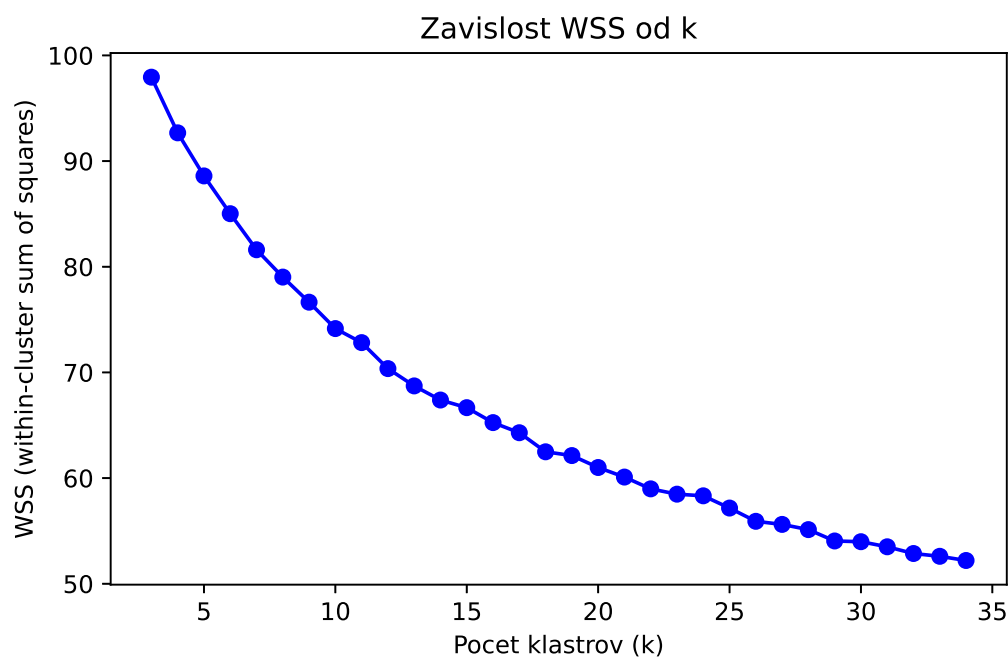
kde θ_m označuje vektor súradníc m -tého centra. Aby algoritmus neskonil v lokálnom minime, bola zvolená viacnásobná inicializácia a výsledok s najnižou vnútrozhlukovou sumou tvorcov WSS.

5 Stanovenie granularity tém

Počet výsledných zhlukov k nebol určený dopredu, ale bol odhadovaný pomocou metódy laka. Sleduje sa pritom pokles hodnoty WSS v závislosti od potu k :

$$\text{WSS} = \sum_{i=1}^n l_{\Theta}^k(x_i) \quad (4)$$

Pri zvyšovaní k hodnota WSS spravidla klesá. Optimálna hodnota k sa zvolí v mieste, kde sa krivka zane postupne vyrovnáva, teda v lakti. Tento bod predstavuje kompromis medzi presnosťou a zloitosťou modelu.



Obr. 1: Závislos sumy tvorcov (WSS) od potu zhlukov k . Lake indikuje optimálnu granularitu tém.

6 Interpretácia objavených tém

Po vykonaní zhlukovania boli centrá jednotlivých zhlukov interpretované ako reprezentanti tém. Pouitá bola inverzná transformácia:

$$x_{ij} \approx \sum_{k=1}^m \phi_{jk} z_{ik} \quad (5)$$

Na základe hodnôt v centre sa vybrali tie výrazy, ktoré mali najvyšie váhy (koeficienty ϕ_{jk}). Tieto slová vystihujú obsah príslušného zhľuku a umoujú kvalitatívne posúdi, aké témy boli v kolekcii textov objavené.

7 Záver

Redukcia rozmerov pomocou PCA zlepila stabilitu a itatenos výsledkov. Zhľukovanie metódou K-means v kombinácii s výpotom WSS poskytlo praktický nástroj na urenie potu tém.

Pre finálny model s $k = 8$ boli identifikované nasledujúce tematické oblasti (zoradené podľa dôleitosti slov v centre):

- Gastronómia: food, city, restaurants, new, york, restaurant, wine.
- Biznis a správy: company, business, said, week, years, year.
- éfkuchári a menu: restaurant, mr, chef, opens, news, bar, menu.
- Literatúra a ivot: life, world, new, book, home, family, father.
- Spolonos a udia: black, white, people, says, men, women, century.
- Film a TV: movie, tv, times, film, series, story, based.
- Dizajn a architektúra: house, designer, design, early, years, century.
- Umenie a práca: work, artist, art, day, young, life.