# Iskanje po zbirki dokumentov

26. marec 2015

**Povzetek**

Izdelajte iskalnik relevantnih dokumentov po ključnih besedah z metodo *latentnega semantičnega indeksiranja* (LSI). Metode, ki izberejo le dokumente, ki vsebujejo natanko iskane besede, so precej nenatančne. Ljudje namreč uporabljamo veliko sopomenk, ki jih preproste metode ne povežejo. Metoda LSI zgradi model, ki združuje več besed v pojme in zato najde tudi dokumente, ki so relevantni, pa ne vsebujejo iskalne besede.

Izdelajte program, ki bo v dani zbirki za dane ključne besede poiskal najbolj relevantne dokumente. Nalogo rešite v več korakih. Glejte tudi [1].

1. Iz zbirke dokumentov zgradite matriko $A$ povezav med besedami in dokumenti. Vsak dokument naj ima v matriki svojo stolpec, vsaka beseda pa svojo vrstico. Element $a_{ij}$ naj bo frekvenca $i$-te besede v $j$-tem dokumentu.

2. Matriko $A$ razcepite z odrezanim SVD razcepom $A = U_k S_k V_k^T$, ki obdrži le $k$ največjih singularnih vrednosti. Razmislite kaj predstavljajo stolpci matrike $U_k$ in matrike $V_k$. Odrezan SVD zmanjša t. i. "overfitting" (preveliko prilagoditev modela podatkom, kar povzroči povečan vpliv šuma).

3. Iskani niz besed (poizvedbo) zapišite z vektorjem $q$. Iz poizvedbe $q$ generirajte vektor v prostoru dokumentov s formulo

$$\hat{q} = q^T U_k S_k^{-1}$$

   Iskanim dokumentom ustrezajo stolpci $V_k$, ki so dovolj blizu vektorju $\hat{q}$. Za razdaljo uporabite kosinus kota med dvema vektorjema in ne Evklidske razdalje med njima. Poizvedba naj vrne dokumente, za katere je kosinus večji od izbrane mejne vrednosti. Preskusite različne mejne vrednosti kosinusa, pri kateri izberemo dokument (0.9, 0.7, 0.6, ...).

4. Metodo je mogoče izboljšati, če frekvence v matriki $A$ nadomestimo z bolj kompleksnimi merami. V splošnem lahko element matrike zapišemo kot produkt

$$a_{ij} = L_{ij} \cdot G_i,$$

   kjer je $L_{ij}$ lokalna mera za pomembnost besede v posameznem dokumentu, $G_i$ pa globalna mera pomembnosti posamezne besede. Preiskusite shemo,

pri kateri je lokalna mera dana z logaritmom frekvence $f_{ij}$ $i$-te besede v $j$-tem dokumentu:

$$L_{ij} = \log(f_{ij} + 1).$$

Globalna mera pa je izračunana s pomočjo entropije

$$G_i = 1 - \sum_j \frac{p_{ij} \log(p_{ij})}{\log n},$$

kjer je $n$ število dokumentov v zbirki,

$$p_{ij} = \frac{f_{ij}}{gf_i}$$

in $gf_i$ frekvenca besede v celotni zbirki. Glejte tudi [2].

5. *Dodajanje dokumentov in besed.* Razmislite, kako bi v model dodali nove dokument ali besede, ne da bi bilo treba ponovno izračunati SVD razcep matrike $A$.

Program preiskusite na podatkih, ki jih dobite npr. na Classic SMART datasets. Lahko pa zudi sami zgradite zbirko iz prosto dostopnih dokumentov.

# Literatura


[1] M. W. Berry, S.T. Dumais, G.W. O'Brien, Michael W. Berry, Susan T. Dumais, and Gavin. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573–595, 1995.

[2] Susan T. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2):229–236, 1991.