Povzetek dokumenta

26. marec 2015

Povzetek

Naloga je izdelati program, ki za dani dokument poišče primeren povzetek. Uporabite metodo *latentnega semantičnega indeksiranja* (LSI) in med povedmni v besedilu poiščete tiste, ki najbolje predstavljajo celoten dokument.

Izdelajte program, ki bo v danem dokumentu poiskal povedi, ki največ povedo o temi dokumenta. Nalogo rešite v več korakih. Glejte tudi [2].

1. Iz dokumenta zgradite matriko A, ki povezuje besede in povedi v dokumentu. Vsaka poved naj ima v matriki svoj stolpec, vsaka beseda pa svojo vrstico. Element a_{ij} naj bo frekvenca i-te besede v j-ti povedi.

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. Za vsako singularno vrednost iz S izberite poved, ki ima največjo ustrezno komponento. Povzetek sestavite iz tako izbranih povedi za nekaj največjih singularnih vrednosti.

4. Stavke za povzetek lahko izberete tudi na podlagi celotne "utežene dolžine", ki upošteva tudi singularne vrednosti s_i:

   $$\|x\|_s = \sqrt{(x_1 s_1)^2 + (x_2 s_2)^2 + \ldots + (x_k s_k)^2}.$$

   Primerjajte povzetek, ki ga dobite na ta način s povzetkom iz prejšnje točke.

5. 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 posamezni povedi, 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-ti povedi:

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

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 povedi v dokumentu,

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

in gf_i frekvenca besede v celotnem dokumentu. Podrobnosti so v [1]. Preverite ali zgoraj opisana mera izboljša kvaliteto abstrakta.

Literatura

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

[2] Josef Steinberger and Karel Jezek. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM'04*, pages 93–100, 2004.