Summary of the document

26. march 2015

Summary

The task is to create a program that finds a suitable summary for a given document. Use the Latent Semantic Indexing (LSI) method to find the sentences in the text that best represent the whole document.

Create a program that will find the sentences in a given document that say the most about the topic of the document. You will solve the task in several steps. See also [2].

1. From the document, build a matrix A that links the words and sentences in the document. Each sentence should have its own column in the matrix and each word its own row. Let element a_{ij} be the frequency of the i-th word in the j-th sentence.

2. Split matrix A with the SVD split cut off $A = U_k S_k V_k^T$, which only holds to the largest singular values. Consider what the columns of the matrix represent U_k and matrix V_k. A truncated SVD reduces the so-called "overfitting" (overfitting of the model to the data, resulting in an increased impact of the sum).

3. For each singular value in S, select the sentence that has the largest corresponding component. Compose a summary from the sentences so selected for the few largest singular values.

4. You can also select sentences for summary based on the total "weighted length", which also takes into account the singular values of s_i :

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

   Compare the summary you get in this way with the summary from the previous section.

5. The method can be improved by replacing the frequencies in matrix A with more complex measures. In general, an element of the matrix can be written as a product of

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

   where L_{ij} a local measure of the importance of a word in a sentence, G_i a global measure of the importance of a particular word. Try the scheme, with

which is a local measure given by the logarithm of the frequency f_{ij} of the i-th word in the j-th sentence:

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

The global measure is expressed in terms of entropy

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

where n is the number of sentences in the document,

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

And gf_i the frequency of the word in the whole document. See [1] for details. Check whether the measure described above improves the quality of the abstract.

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.