Multimedia Information Extraction and Retrieval SoSe 2010
Exercise Sheet 3

Prof. Dr. Ralf Möller, Sebastian Wandelt

05.05.2010

1. Explain the *bag of words* view of documents. Name examples for areas which use that bag of words view.

   Solution:

   Latent Semantic Analysis, Bayesian Spam Filtering (two bags: spam and ham, determine probability from which bag), Other Bayesian Models

2. Explain the difference between term *frequency* and *count*.

   Solution:

   In a lot of IR literature, frequency is used to mean count. Thus term frequency in IR literature is used to mean number of occurrences in a doc. Not divided by document length (which would actually make it a frequency)

3. How to represent documents in a vector space? How big is that vector space potentially? What problems occur? **Solution:**

   Even with stemming more than 50000 axes, slow performance

4. How can we reduce the number of axes in the document vector space?

   Solution:

   Approximation of matrizes (vector spaces, LSI).

5. How can we define similarity of documents in the vector model? What are the advantages and disadvantages of these approaches? **Solution:**

   Documents close to each other are similar.

   - *distance*
   - *cosinus similarity*

6. Determine the similarity of the following texts with respect to the similarity measures in the last task.

   - „LSI tutorials and fast tracks.“
   - „Books on semantic analysis.“
   - „Learning latent semantic indexing.“
   - „Advances in structures and advances in indexing.“
   - „Analysis of latent structures.“

   Solution:

   see exercise session

7. Show that, for normalized vectors, Euclidean distance gives the same proximity ordering as the cosine measure.

   Solution:

   Simple transformation ... see lecture slides.