Multimedia Information Extraction and Retrieval
Summer Term 2012
Exercise Sheet 3

Ralf Möller, Karsten Martiny

Exercise Session:
Thursday, May 3, 2012, 8.00-8.45, SBS95-D1025

1. What is the advantage of using a logarithmic term frequency measure instead of a plain linear tf?
2. Why is the idf of a term always finite?
3. What is the idf of a term that occurs in every document? Compare this with the use of stop word lists.
4. Explain the idea of using tf.idf weights for scoring.
5. Can the tf-idf weight of a term in a document exceed 1?
6. How can one represent documents in a vector space? How big is that vector space potentially? What problems occur?
7. A vector space model can be used for determining similiarity of documents by assuming that documents that are "closer together" have the same topic. Which properties should hold for a meaningful proximity unit?
8. Explain how a vector space model can be used to answer bag-of-words queries.
9. Determine the similarity of each the following texts(=documents) with respect to the cosine similarity measure. It is sufficient if you write down the formula for each pair of documents. You don't need to compute the actual value.
   - "Hello you"
   - "Where are you?"
   - "Are you feeling well?"