Information Retrieval, T-75.4400
Exam 1.9.2014

Answer to all four following questions. The supplementary material is given in the end of this exam. The questions are available in English. You may answer in English, Finnish or Swedish.

1. Explain briefly the following concepts and give an example of the concept in practical information retrieval setting. (6p)
   a. Relevance feedback
   b. Phrase query
   c. Inverted index
   d. Precision
   e. Query expansion
   f. Stemming

2. The end of this exam contains an example dataset. Do the following (2p+2p+2p):
   a. Construct a term-document matrix using the example dataset.
   b. Use a unigram language model without smoothing and prior information and compute a probability for documents for a query consisting of the following terms: "cats", "play", "piano" using a maximum likelihood estimate.
   c. Compute the same probabilities as in the previous task (b), but with a language model using mixture model smoothing with $\lambda = 0.5$. Recall that the mixture model can be computed using the following formula:

$$P(q|d) \propto \prod_{1 \leq k \leq |q|} (\lambda P(t_k|M_d) + (1-\lambda)P(t_k|M_c))$$

3. Explain what the following psychological phenomena mean and how they can affect search user-interface design. Give examples of search user interface techniques that are related to these phenomena (3p+3p):
   a. Recognition and recall
   b. Anchoring

4. Figure 1. Presents a web link graph. Do the following (2p+2p+2p):
   a. Construct a transition probability matrix for the documents d0…d6 in the web link graph and modify the matrix to utilize teleporting. Use a teleportation rate of 0.14
   b. Compute one iteration (after initialization) of PageRank for each document using the power iteration method. Initialize the vector for iteration 0 using a uniform distribution, i.e. probability of 1/7 for each document. Use the matrix with teleportation resulting form the previous phase (a). To keep the manual computation feasible, the matrix resulting from the previous phase can be rounded up to a 2 decimals.
   c. Shortly explain how PageRank can be used in Web search systems.

Supplementary materials:

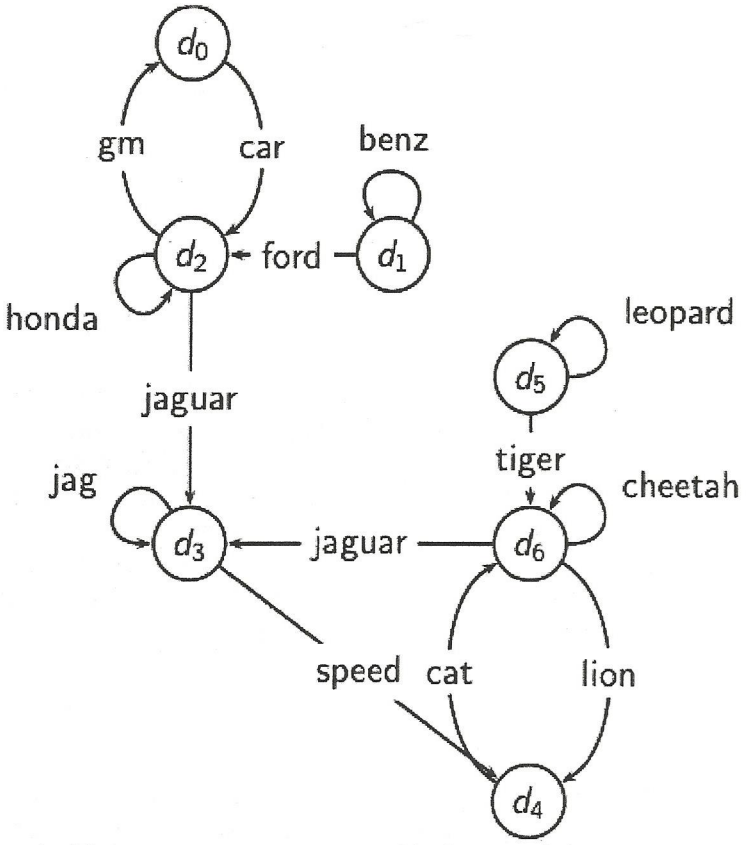

Figure 1. A Web link graph. The arcs are annotated with the word that occurs in the anchor text of the corresponding link.

Example dataset (bolded words are the words occurring in each of the documents d0…d2):

d0: ***Cats play piano on YouTube***

d1: ***Dogs cannot play piano***

d2: ***Cats and dogs are animals***