Information Retrieval, T-75.4400
Exam 23.5.2013

Answer to all four following questions. The supplementary material is given in the end of this exam containing a set of documents that you need to use in answering to question, and a link graph that you need to use in answering to question 4. The questions are available in English, but you may answer in English, Finnish or Swedish. See also the back of this paper!

1. Explain briefly the following concepts. Give an example if requested. (6p)
   a. Ad-hoc information retrieval
   b. Anchoring effect (in query reformulation)
   c. Inverted index. Give an example.
   d. Term relevance feedback
   e. Keywords-in-context (KWIC)
   f. Query expansion

2. The end of this exam contains an example dataset. Do the following (3p+3p):
   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 the following (3p+3p):
   a. What do "semantic constructs" of natural language refer to? What kind of problems these can cause in information retrieval? Give examples of at least two typical problems.
   b. Explain the basic principles of Latent Semantic Indexing (LSI). Give an example on how LSI can be useful in information retrieval.

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 decimal places.
   c. Shortly explain what PageRank (or similar) measure can be used for in information retrieval systems. How anchor text present in the the Web graph could be utilized? Give an example.

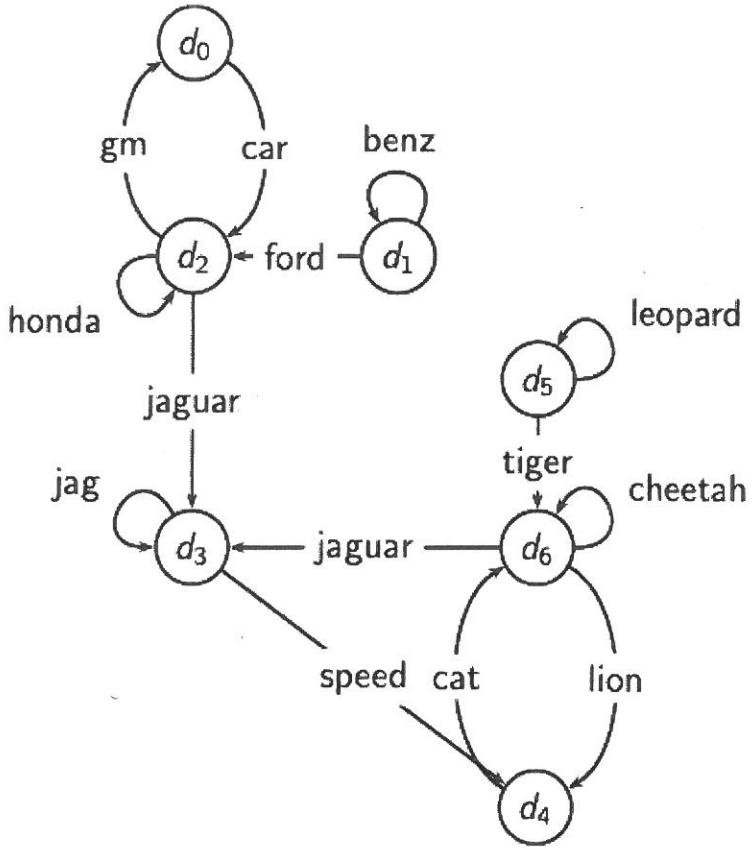

Figure 1. A small web link graph. The arcs are annotated with the word that occurs in the anchor text of the corresponding link.

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***