

Internet Information 2006-2007 (MIKII6)

Mid-term exam

March 26, 2007 14:00–17:00, room A.404

This is not an open book exam. No use of additional materials is allowed. Mobile phones and other communication devices should be switched off and stored out of sight. You can bring a calculator the the exam. You can answer questions in English or Dutch.

There are three sets of questions, each should take you about 45 minutes, but you have three hours in total. Each question set accounts for 1/3 of the mark. Within a set, questions have equal weights.

Part 1

Answers to the questions in Part 1 usually consist of 2–5 sentences. Provide formulas where appropriate, clearly explaining your notation and terminology.

1. Why is stemming typically not used in Web IR?
2. Describe the structure of the inverted index for document retrieval based on vector space model.
3. What is blind relevance feedback?
4. Does vector space models use smoothing? If yes, describe one such smoothing method; if not, explain why.
5. What is multimodal search? Give an example where multimodal search is necessary.
6. What is the purpose of the PageRank and HITS algorithms? How are they similar and how they differ
7. Why *inverted document frequency*, rather than *document frequency* is typically used in some IR models?
8. How Named Entity Recognition can be modeled as a labeling task?
9. Describe the basic architecture of a web crawler.
10. Give examples of evaluation measures appropriate for (a) Question Answering and (b) ad-hoc document retrieval.

Part 2

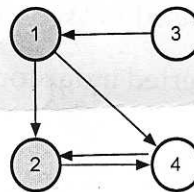
Questions in Part 2 are about computing the required values. In addition to the numerical results, you should provide the formulas you use and explain your notation.

1. Calculate the exact value of the Jaccard similarity based on 3-word shingles for the following two documents:

- $D_1 = \text{advantages of Jelinek-Mercer smoothing over Dirichlet smoothing}$
- $D_2 = \text{some advantages of Dirichlet smoothing over Jelinek-Mercer}$

Calculate an approximation of the Jaccard similarity using three arbitrary permutations.

2. You need to calculate hubs and authority scores for the HITS algorithm. In the graph below, gray nodes represent pages in the root set (returned by a retrieval algorithm) and white nodes represent the pages of the base set added after the expansion. Starting from uniform scores, calculate first two iterations of the HITS algorithm.



Part 3

The questions in this part are open-ended; they do not have single correct answers. We expect well-motivated design choices and clear explanations of approaches you think are most appropriate, including formulas with clearly defined notation, where appropriate. Answers are typically between half a page and one page long.

1. You are asked to design a search system for a large discussion forum, that given a topic, would allow a user to find recent important developments on the topic. What are the aspects of this data collection that can improve basic search? Describe a retrieval model that would use the structure of the collection to implement an effective and convenient search system.
2. You need to design a language modeling-based retrieval system for a large audio collection of monolingual news broadcasts. You can use a near-perfect audio segmentation software and an advanced speech recognition system that is capable of finding N most probable transcripts of a segment, along with their probabilities. Design a LM retrieval model that incorporates these values.