# Part II Information Retrieval Exercises

## Lectures 1-4

Lecture 1 is an introduction to IR, and especially the vector space model for the document retrieval problem. Lecture 2 briefly introduces the Boolean model of document retrieval and then gives more detail for the vector-space model. It finishes by talking about indexing and representational issues, such as stemming. Lecture 3 discusses evaluation in IR, especially the TREC competitions and evaluation for the document retrieval task. Lecture 4 is primarily a discussion of PageRank.

1. Small practical test.

- Build a term-document matrix for three short documents of your choice.
- Weight terms by presence/absence (binary) and by TF*IDF (with estimated IDFs).
- Write a suitable query, and calculate document-query similarity, using:
  - cosine
  - inner product (i.e. cosine without normalisation)

a. Does the normalisation have any effect?

2. This question is about the methods used by various search engines. By posing queries to a search engine, try and answer the following questions:

- Does it use a stop list? (Can you search for the phrase “The The”?)
- Does it use stemming?
- Does it normalise words to lower case?
- Does it use phrases (in addition to the quoting mechanism used by Google)?
- Does it use any additional “linguistically motivated” processing?

Do the answers to these questions vary across search engines? Compare your answers for three major search engines, e.g. Google, Yahoo, MSN search.

3. Knowing the sense (meaning) of a query term may help a document retrieval system, especially for short queries. Why does knowing the senses of query terms become less useful for longer queries?

4. Boolean model and stemming.

Comment on the following statements:

a. *Stemming never lowers precision of a Boolean retrieval system.*

b. *Stemming never lowers recall of a Boolean retrieval system.*

HINT: think about the NOT operator.

5. Evaluation metrics.

a. What is the difference between *accuracy*, and *precision* and *recall*?

b. Form a Google query for the information need on slide 10 of lecture 3. What is the precision of the Google search for the first 10, 20, 30 documents?

c. Why is calculating recall a problem for document retrieval? Explain the TREC solution (pooling).

6. Web search and PageRank.

a. Consider a web graph with three nodes 1, 2 and 3. The links are as follows: $1 \rightarrow 2, 3 \rightarrow 2, 2 \rightarrow 1, 2 \rightarrow 3$. Write down the transition probability matrices for the surfer's walk with teleporting, for the following three values of the teleport probability: $\alpha = 0$; $\alpha = 0.5$; $\alpha = 1$.

b. A user of a browser can, in addition to clicking a hyperlink on the page $x$ he is browsing, use the *back button* to go back to the page from which he arrived at $x$. Can such a user of back buttons be modelled as a Markov chain? If not, why not?

c. Explain why the page rank of every web page is at least $\alpha/N$, where $\alpha$ is the teleport probability and $N$ is the total number of web pages. (A formal proof is not required.) What does this imply about the page rank values as $\alpha$ approaches 1? Explain.

d. Describe some of the properties of the Web which make document retrieval on the Web challenging. What strategies do search engines use to overcome these challenges?

SCC November 2009