Resit examination: Information retrieval 2011

10 Februari 2011.

Please answer the following 21 questions as completely and concisely as possible. Good luck.

Question 1 (5 points): In which order can you best process the following boolean search query for fast search results, given the inverted indices below:

Query: information AND retrieval AND Friday

Inverted index:

Information -> 3,5,7,9,11,12,13
RoomP647 -> 3,4,5,6,8,9
retrieval -> 3,4,5,11,12
Friday -> 3,4,5,8,9,14,15,16

Question 2 (5 points): Which of the following document sets pose the biggest problem for the evaluation of search engines? Explain your answer.

True Negatives

True Positives

False Positives

False Negatives

Question 3 (10 points): Explain the Sign test in your own words, and what you determine with it?

Question 4 (10 points): Create the inverted index for the following documents:

1) my son and I
2) like teacher and father
3) like father and son
4) I like my son

Question 5 (10 poins): Explain how the following three cases can happen:

- stemming increases Precision
- stemming decreases Precision
- stemming increases Recall

Question 6 (5 points): What are the two greatest shortcomings of the boolean retrieval model?

Question 7 (5 points): What is the main reason why most vector retrieval models use inverted document frequency in favor of inverted collection frequency?

Questions 8 (5 points): Given a document that contains 4 times the word "information". It occurs twice in the query, which is of length 5. Consider now a collection of 10.000 documents, containing 400 times the term "information". In total the word occurs in 50 documents. The average length of each document is 43, and there are in total 450000 words in the corpus.

What is the tf-idf value (according to the definition given in the lecture, and repeated above) of the word "information". If you do not have a caculator, just fill in the formula.

Question 9 (10 points): Explain the cosine measure, and its function in Vector-Space retrieval?

Question 10 (15 points): Explain Heap's and Zipf's laws, and why Heaps' law follows from Zipf's.

Question 11 (15 points): Let X be a document-term matrix and USV^T its singular value decomposed form, such that X=USV^T. Explain what is in U, S and V.

Question 12 (10 points): What is topic drift?

Question 13 (5 points): What do you get when you multiply a document-term matrix by itself?

Question 14 (5 points): Which of the following probabilities can be expected to be the LOWEST in an average english text? Explain why.

P(retrieval|information)
P(information|retrieval)

Question 15 (5 points): Calculate the centroid of the following document vectors:

d1=(4,3,1),
d2=(5,10,2),
d3=(7,7,3),
d4=(12,5,4)

Question 16 (5 points) Can the unigram model account for word order? Explain your answer.

Question 17 (15 points): Explain PageRank and HITs, and the difference between the two.

Question 18 (10 points): In order to improve the quality of a classifier for text categorisation you have to choose a suitable set of attributes (words). Explain the difference between using a stop-word list on the one hand and automatic attribute selection algorithms, eg. based on Information Gain, as offered in Weka?

Question 19 (10 points): Explain why accuracy is a rather useless measure in text-classification. Why is the f-measure more suitable?

Question 20 (5 points): Describe what is meant by the bow-shape of the Web, and its effect on crawling?

Question 21 (10 points): in 90% of all the documents about Dutch politics contain the term "Wilders". In all the documents on the Dutch web, 1 out of 100.000 mention "Wilders". About 0.1% of all documents on the Dutch web are about Dutch politics.

Use Bayes rule to to calculate the probability that a document containing the term "Wilders" is about Dutch politics.

END OF THE EXAM