Web Information Retrieval

Exam
April 16th, 2015
Time available: 90 minuti
5 points for each problem

Problem 1

1. Describe what is *document ranking* in the vectorial model with respect to a query `q` when using cosine scoring. How are documents and queries represented? Describe how the postings lists are adapted to implement cosine scoring.
2. Define the tf-idf weight of a term, also describing the tf and idf components and their meanings.
3. Describe the pseudocode of an exact Top-K algorithm in the vectorial model.

Problem 2

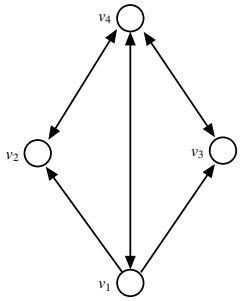
The following list of R's and N's represents relevant (R) and nonrelevant (N) returned documents in a ranked list of a collection of 30 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list.

R R N R N N R R N N N R N N N N R N N R N N N N N N R N R N

1. What are precision and recall of the system on the top 10?
2. Draw the precision-recall curve.
3. Must there always be a break-even point between precision and recall? Either show there must be or give a counter-example.

Problem 3

1. We are given the following graph. Write down all the necessary equations needed to calculate the pagerank, for a general teleporting probability α.



2. Assume that teleporting is modified as follows: at every step, wtih probability α, the random surfer jumps to vertex v_1. Is the resulting process still a Markov chain? Either prove or use a counterexample to disprove.

Problem 4

1. Describe the assumptions of a Naive Bayes classifier in the `bag of words` model.
2. Compute the coefficients of a boolean classifier without `smoothing` on the following 4 training documents:

   (a) `browsing tiger safari.  apple`

(b) `africa video lion.  not apple`

(c) `lion mountain osx.  apple`

(d) `mountain safari browsing tiger.  not apple`

3. Classify the query document: `lion mountain safari`

I consent to publication of the results of the exam on the Web

Firstname and Lastname in block letters.....

Signature