Nome e Cognome:

Matricola:

Web Information Retrieval

Exam, 6 July 2013, *Time available: 100 minuti*
4 points/problem

Problema 1

1. How should the Boolean query `x AND NOT z` be handled? Why is naive evaluation of this query normally very expensive? Write out a postings merge algorithm that evaluates this query efficiently.
2. Describe what structure we need to be able to answer phrase queries such as: "In un piatto poco cupo, poco pepe cape." Give an example of such an index by constructing some sample documents and presenting the corresponding index.

Problema 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 is the precision of the system on the top 20?
2. What is the recall on the top 20?
3. Draw the precision-recall curve.

Problema 3

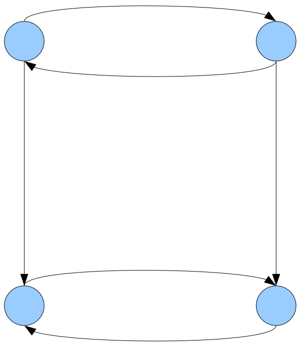
1. Write the *tf* × *idf* weighting equation. Explain what each term represents, and the reasoning about the equation.
2. Consider an IR system where we use the *tf* × *idf* weighting scheme. We compare three pairs of documents:
   (a) Two docs that have only frequent words (the, a, an, of, etc.) in common.
   (b) Two docs that have no word in common.
   (c) Two docs that have many rare words in common.

   Rank the above pairs according to their cosine similarity score and explain convincingly the reasoning behind your choice.
3. State two reasons that in IR we usually use cosine similarity instead of Euclidean distance.

Problema 4

1. What is the importance of the teleporting probability with respect to the convergence of pagerank?
2. We are given the following graph. Write down all the necessary equations needed to calculate the pagerank, for a general teleporting probability α.

3. Compute the pagerank of each node for teleporting probability $\alpha = 1/2$.



Problema 5

1. Describe the assumptions of a Naive Bayes Bernoulli classifier.
2. Compute the coefficients of a boolean classifier on the following 4 training documents:
   (a) `safari jeep lion. not car`
   (b) `lion toyota safari. not car`
   (c) `jeep jaguar. car`
   (d) `toyota wolkswagen jeep. car`
3. Classify the query document: `toyota jeep`

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

Firstname and Lastname in block letters...

Signature