Nome e Cognome:

Matricola:

Web Information Retrieval

Exam, 11 Settembre 2013, *Time available: 100 minuti*
4 points/problem

Problema 1

1. Give the pseudo-code of a linear-time algorithm for the intersection of three posting lists.
2. Consider the query `Web AND Information AND Retrieval`:

   ```
   Web   [5; 7; 12; 19; 25]

   Information   [5; 8; 12; 19]

   Retrieval   [8; 12; 19; 25]
   ```

   Work out how many comparisons would be done to intersect the three postings lists.
3. Modify the algorithm to consider queries of the type `Web AND Information AND NOT Retrieval`.

Problema 2

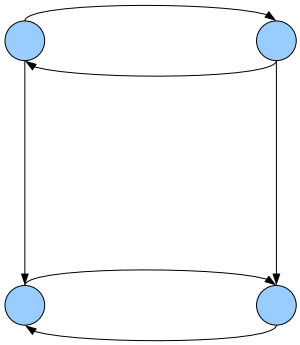
1. Show how we can compress the list `[10, 30 , 40 , 41, 55, 75 , 78, 105]` using
   (a) Variable byte encoding.
   (b) γ encoding.
2. Show how to decompress online the compressed list.

Problema 3

1. Show that for normalized vectors, Euclidean distance gives the same proximity ordering as the cosine measure.
2. Given a query vector, show how to compute efficiently the top k nearest documents according to cosine similarity.

Problema 4

1. We are given the following graph. Compute the pagerank score of each node for teleporting probability $\alpha = 0$.
2. Compute the pagerank score of each node for $\alpha = 1/2$.



Problema 5

1. Explain briefly how the k-means algorithm works. Write the algorithm.
2. You are given the following example. Show that if the initial cluster assignment is unlucky the k-means solution might be bad.

3. Explain briefly why the k-means algorithm converges.

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

Firstname and Lastname in block letters.....

Signature