Nome e Cognome:

Matricola:

Ricerca dell'Informazione nel Web

Compito di esame, *tempo a disposizione: 90 minuti*
5 punti/problema

Problema 1

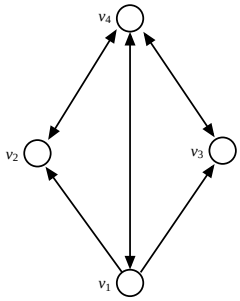
1. Are the following statements true or false? Briefly explain your answers.
   (a) In a Boolean retrieval system, stemming never lowers precision.
   (b) In a Boolean retrieval system, stemming never lowers recall.
   (c) Stemming increases the size of the vocabulary.
   (d) Stemming should be invoked at indexing time but not while processing a query.
2. Why are skip pointers not useful for queries of the form `x OR y`?
3. Assume a biword index. Give an example of a document which will be returned for a query of `New York University` but is actually a false positive which should not be returned.

Problema 2

1. Show that for normalized vectors, Euclidean distance gives the same proximity ordering as the cosine measure.
2. Is this true for non-normalized vectors? Prove or disprove.

Problema 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. Compute the pagerank of each node for teleporting probability $\alpha = 1/2$.
3. Prove that for any graph the pagerank of each node is at least α/N.



Problema 4

1. Explain briefly how the k-NN algorithm works.
2. Explain briefly how we can perform k-NN classification in the vector-space model using the inverted-list data structure.

3. Explain briefly how the k-means algorithm works. Write the algorithm.

4. You are given the following example. Show that if the initial cluster assignment is unlucky the k-means solution might be bad.