Nome e Cognome:

Matricola:

Ricerca dell'Informazione nel Web

Compito di esame del 20 Febbraio 2013, *tempo a disposizione: 90 minuti*
5 punti/problema

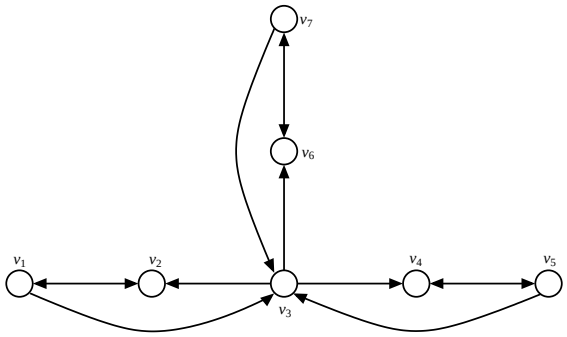
Problema 1

Show how we can compress the list `[10, 20, 25, 33, 40, 50, 72, 92]` using

1. Variable byte encoding.
2. γ encoding.

Problema 2

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 3

1. Write the *tf*$\times$ *idf* weighting equation. Explain what each term represents, and the reasoning about the equation.
2. Consider an IR system where we use the *tf*$\times$ *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. 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.

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

Firstname and Lastname in block letters.....

Signature