**Information Retrieval**

**15 June 2018**

**Ex 1 [rank 3+3+3]** Let you be given four texts

D1= “a cose fatte”

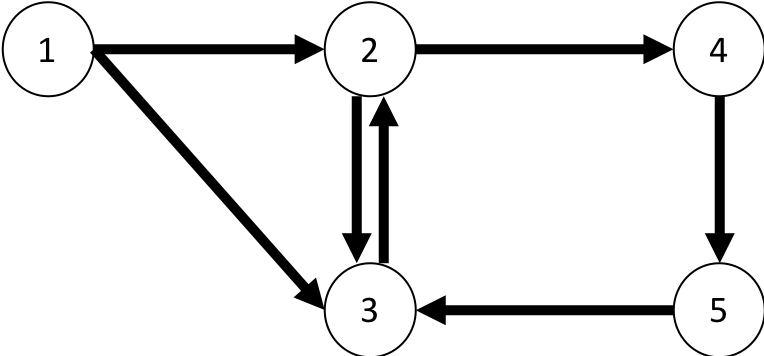
D2= “cosa fatta, cose fatte”

D3= “capo ha”

D4= “cosa cosa cosa a capo capo”

1. Show the inverted index built on these four documents
2. Show the TF-IDF vectors for these documents, by assuming that the logarithm is in base 2.
3. Compute the document which is more similar to the query [cosa capo], by using the cosine similarity without dividing by the norms of the vectors.

**Ex 2 [rank 5]** Let you be given the graph:



Show the execution of two steps of PageRank, by assuming uniform initial probability distribution and uniform teleportation step, with alpha = ½.

**Ex 3 [punti 3+3+3]** Answer to the following questions:

* Describe the Rocchio’s method.
* State the Johnson-Linderstrauss lemma and comment on its use.
* Define the features: link probability of an anchor, and commonness of a Wikipedia page wrt to an anchor, as used by TagMe.

**Ex 4 [punti 3+4]** Given the set of strings S = {abito, bit, kino, kit, skin, skinny}.

* Build a 2-gram index for S
* Show how the 2-gram index can be used to speed-up the execution of a query for the string “bikini” in S with edit distance e=2.