Information Retrieval – exercises
05 June 2023 – time 60 minutes

Name and Surname: #matricola:

Question #1 [rank 4]. You are given the two files:

F_old = “AAAA BBBB”, F_new = “A BBBB BA”,

and assume a block size B=3 chars (SPACE is a char).

- Show the execution of the algorithm zsync. *(comment the various steps)*

Question #2 [rank 3+3]. Given the set of strings S={aba, abc, baac, babc}.

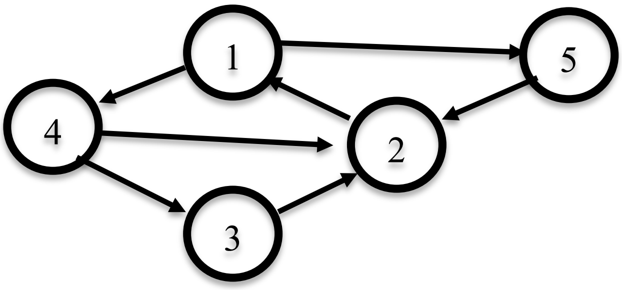
- Show the (compacted) trie T built on S
- Show how to search for the lexicographic position of “abb”

Question #3 [rank 2+3+2]. Let you be given 3 documents:

D1= “A NICE THING”
D2= “THING DONE, THINGS DONE”
D3= “THING THING THING DONE DONE”

a) Show the inverted index built on these 3 documents;
b) Show the TF-IDF vectors for these documents, by assuming that the logarithm is in base 2 (*hint:* you can keep the LOG-formula as they are);
c) Compute the document which is more similar to the query [NICE THING], by using the cosine similarity without dividing by the norms of the vectors.

Question #4 [rank 3]. Given the graph



Compute one step of Personalized PageRank (PPR) with respect to the set S = {1, 2}, by assuming a uniform starting distribution and parameter alpha=0.5.

Information Retrieval – theory
5 June 2023 – time 45 minutes

Name and Surname: **#matricola:**

Question #1 [scores 3] Show and comment how to efficiently compute the Hamming distance between pairs of binary vectors by using the Locality Sensitive Hashing approach.

Question #2 [rank 3] Show how to compute text summarization by using a graph and Page Rank.

Question #3 [rank 2+2] Define what it is a wild-card query, and show how to solve it via a Permuterm index.