Information Retrieval – EXERCISES
4 September 2024 – time 60 minutes

Name and Surname: #matricola:

Question #1 [scores 5] Given 4 strings S = { abaco, basco, raco, vasto }, describe how Z-delta compresses these files via a properly constructed weighted directed graph.

Question #2 [rank 2+3] Given the dictionary of strings D = {abba, abc, babb} construct a bigram index (hence k=2). Then given the string Q = "abcc" use the overlap distance to filter a set of strings from D that are potential candidate for an edit distance e=1.

Question #3 [scores 4] Given the four texts:

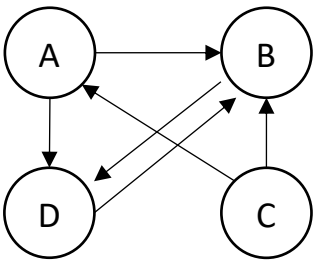
T_1 = "I have a pen"
T_2 = "I have a pineapple"
T_3 = "a pen a pineapple"
T_4 = "a pen an apple pen"

by using the TF-IDF vectors of the four texts above (logs are in base two), find the most similar text to the query "have a apple" (do not apply any normalization to the cosine score).

Question #4 [scores 2+2+1] Given the sequence of integers S = (3, 4, 12, 20, 24, 27, 28, 29, 31) show how to compress:

a) S via the Elias-Fano code.
b) the gap-encoded S via the gamma code.
c) the gap-encoded S via the PForDelta code with base = 1 and b = 2.

Question #5 [scores 3] Compute the authority score and the hub score of nodes A and B in the following graph via one step of the HITS algorithm. Assume that the starting vectors of authority and hub scores are both equal to [2, 1, 2, 1].



Information Retrieval – THEORY
4 September 2024 – time 45 minutes

Name and Surname: #matricola:

Question #1 [rank 2+2] Describe the Front queue and the Back queue in the Mercator crawler, and state/comment their goals.

Question #2 [rank 2] Describe the rule that allows Block-Max WAND to skip scoring the documents in a block.

Question #3 [rank 2] Write the PageRank formula, and describe the consequences of setting the parameter α to 0 and to 1.