Information Retrieval – EXERCISES
21 June 2024 – time 60 minutes

Name and Surname: #matricola:

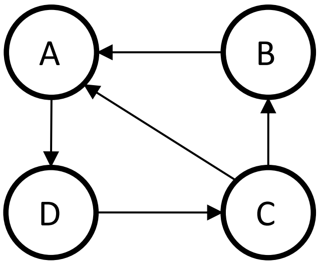
Question #1 [rank 2+2]. Compute the Permuterm index for the two strings {BOSS, POS} and show how it is possible to search for B*S into it.

Question #2 [rank 3]. Given the sets A = { 2, 3, 5, 7 } and B = { 1, 5 } and C = {5, 8, 9}, simulate the min-hashing technique in approximating the Jaccard similarity among them, by using the following three permutations $P_1(x)$ = x mod 11; $P_2(x)$ = 2*x mod 11; $P_3(x)$ = 3*x mod 11.

Question #3 [scores 2+2+2+2] Given the sequence of integers S = (3, 5, 8, 10, 11, 18, 24, 28) 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.
d) the gap-encoded S via t-nibble with t = 3.

Question #4 [scores 1+3+3] Given the following graph:



a) Comment on whether a random walk computed over this graph converges to a single state that is independent of the starting distribution.
b) Compute one step of Personalized PageRank with respect to nodes A and C by assuming a uniform starting probability distribution and α = ½.
c) Compute the authority score and the hub score of nodes A and C 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 [1, 2, 1, 2].

Information Retrieval – THEORY
21 June 2024 – time 45 minutes

Name and Surname: #matricola:

Question #1 [scores 3] Describe the use of the cluster pruning approach over a collection of documents for implementing the approximate top-K retrieval with respect to a query document d.

Question #2 [scores 3] Describe the Rocchio approach to relevance feedback, and comment on its limitations.

Question #3 [scores 2] Write and comment on the Discounted Cumulative Gain (DCG) formula.