Information Retrieval – EXERCISES
11 July 2024 – time 60 minutes

Name and Surname: #matricola:

Question #1 [scores 3+3] Given the dictionary of strings Dict = { CCB, CBCA, ADA }

a) construct a bigram index (hence k=2).
b) Then given the string Q = "BCCB" use the overlap distance to filter a set of strings from Dict that are potential candidate for an edit distance e=1.

Question #2 [scores 3+4] Given the two files

F_{old} = "How_much_is_good", F_{new} = "How_much_are_good",

and a block size B=3 chars (*hint:* if the length is not a multiple of B, add NULL chars).

a) Describe rsync running on them;
b) Describe zsync running on them.

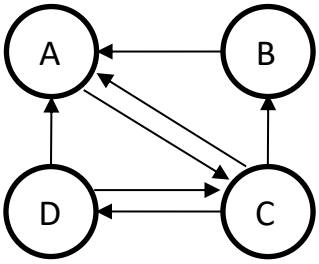
Question #3 [scores 2+2+1+1] Consider the WAND algorithm for examining the head of the following four posting lists:

t_1 → 2, 5, 6, 8, 10, 11, 13
t_2 → 8, 11, 12, 13, 15, 17, 19, 21, 25
t_3 → 5, 6, 7, 8, 10, 12, 13, 21
t_4 → 3, 5, 8, 11, 13, 19, 22

The current threshold is θ = 2.5, and the upper bounds of the scores in each posting list are: $ub_1 = 0.5$, $ub_2 = 0.7$, $ub_3 = 0.8$, $ub_4 = 1$.

a) Which is the candidate docID, and is its full score computed?
b) Suppose instead the algorithm is Blocked-WAND with blocks of size 5 and local upper bounds of the first block in each list equal to $lb_1 = 0.5$, $lb_2 = 0.5$, $lb_3 = 0.4$, $lb_4 = 0.8$. Which is the candidate docID, and its full score is computed?
c) Still considering the Blocked-WAND algorithm and the setting of point b) above, which blocks can be discarded to go to the next docID?
d) Can the value of θ change after the Blocked-WAND step of point b) above? Why?

Question #4 [scores 3] Compute one step of PageRank on the following graph by assuming α = 2/3 and starting probability distribution equal to [1/4, 0, 2/4, 1/4].



Information Retrieval – THEORY
11 July 2024 – time 60 minutes

Name and Surname: #matricola:

Question #1 [scores 2+2]

- Describe the algorithm that computes the LSH-sketch of a binary vector for the case of hamming similarity, and show how it is used to declare that two vectors are “similar”.
- State and prove what is the probability that the above algorithm declares that two vectors are “similar” provided that their real similarity is s.

Question #2 [scores 2] Write the tf-idf formula, and describe what information must be stored in the inverted index to compute it efficiently in time and in space.

Question #3 [scores 2] Describe how Hyperlink-Induced Topic Search (HITS) determines the so-called “root set” and “base set” of Web pages starting from a user query.