Information Retrieval (Final Term)
15 December 2023 – time 1:45 hours

Name and Surname: #matricola:

Question #1 [scores 1+2+2] Given the following four documents:

D_1 = “it was the best of times”
D_2 = “the worst of times”
D_3 = “it was the age of wisdom”
D_4 = “the age of foolishness”

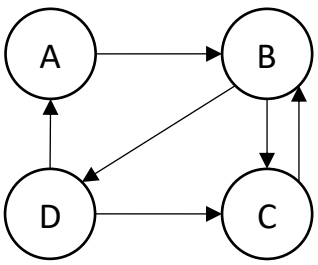
a) Show the inverted index built on these documents.
b) Show the TF-IDF vectors of these documents by assuming logarithms to the base 2, and by not evaluating the logarithms numerically.
c) Find the document that is most similar to the query q = “best age times” by using the dot product (that is, the cosine similarity without normalization).

Question #2 [scores 2+3+2] Given the sorted sequence of integers

S = (1, 4, 5, 10, 16, 19, 23)

a) Show how to compress the gaps between consecutive integers in S via the gamma code.
b) Show how to compress S via the Elias-Fano code.
c) Show how to compress the gaps between consecutive integers in S via the PForDelta code with base = 1 and b = 2.

Question #3 [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 PageRank by assuming a uniform starting probability distribution and α = ½.
c) Compute one step of Personalized PageRank with respect to node D by assuming a uniform starting probability distribution and α = ½.

Question #4 [scores 3+3+1] Consider the WAND algorithm for examining the head of the following four posting lists:

$t_1 \rightarrow 3, 4, 5, 6, 7, 20, 22$
$t_2 \rightarrow 1, 5, 7, 10, 21$
$t_3 \rightarrow 5, 7, 11, 20, 22$
$t_4 \rightarrow 7, 8, 10, 11, 14$

The current threshold is $\theta = 3.3$, and the upper bounds of the scores in each posting list are: $ub_1 = 1$, $ub_2 = 2$, $ub_3 = 0.5$, $ub_4 = 1.2$.

a) Which is the candidate docID, and is its full score computed?
b) Suppose instead the algorithm is Blocked-WAND with blocks of size 3 and local upper bounds of the first block in each list equal to $lb_1 = 1$, $lb_2 = 1.8$, $lb_3 = 0.4$, $lb_4 = 0.8$. 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 block is discarded to go to the next docID?

Question #5 [scores 2] Describe the cluster pruning approach for approximate top-K retrieval.

Question #6 [scores 2] State which are the scores computed by HITS and comment on them briefly.