Information Retrieval

15 December 2020 – time 45 minutes

Question #1 [rank 4+4]. Given the sequence of integers S = (1, 2, 3, 10, 12, 13) describe how you'd compress it via:

- Gamma coding
- PForDelta coding, with b=2 and base=1

Question #2 [rank 5+3]. You are given the following three texts:

a. T1="white book"
b. T2="white book really white"
c. T3="a black book"

- Compute the TF-IDF vectors of the three texts above (logs are in base two, and do not compute them via a calculator).
- Find the most similar text to the query "black book" in the vector space model (no normalization). Assume that idf for the terms of the query are the same as for the terms in the texts.

Question #3 [rank 3+4+2]. Consider the WAND algorithm over the following four posting lists by assuming that at some step the scanning algorithm is examining

t1 → (…, 3, 5, 10, 12)
t2 → (…, 2, 3, 5, 7, 9, 11, 12, 13)
t3 → (…, 1, 5, 13, 20)
t4 → (…, 6, 7, 8, 9, 12, 20)

At that time the current threshold equals 3.5, and the upper bounds of the scores in each posting list are: ub_1 = 1, ub_2 = 2, ub_3 = 1, ub_4 = 2.3.

- Choose the pivot document and discuss if its full score is or it is not evaluated.
- Make another step of the WAND algorithm, choose the next pivot document and discuss what happens to the evaluation of its full score.
- Assume that the above lists are partitioned into blocks of three docIDs each, and assume that the first block of each list has a local upper bound ub^L equal to: ub^L_1 = 1, ub^L_2 = 1, ub^L_3 = 1, ub^L_4 = 2.3. Discuss the execution of the algorithm Blocked-WAND applied to the pivot document chosen as answer of the previous question.

Question #4 [rank 5]. Compute one step of PageRank by assuming a uniform starting distribution, and setting alpha=0.5