Information Retrieval

Final term – 17 January 2022 – time 80 minutes

Question #1 [rank 5+3+3].

- Describe the execution of Rsync that updates the old file F_old = 00111010 available at the client with the new file F_new=001101010 available at the server. Assume that the block size is b=4, and that the hash function is h(x) = x mod 4 where x is the value of a block of 4 bits. Comment on the correctness of the execution.
- Describe the execution of Zsync that updates the old file F_old = 00111000 available at the client with the new file F_new=00111111 available at the server, under the following scenarios:
    - **Case i:** Assume that the block size is b=4 bits, and that the hash function is h(x) = x mod 4 where x is the integer value represented by a block of 4 bits. Comment on the correctness of the execution.
    - **Case ii:** Assume that the block size is b=4 bits, and that the hash function is h(x) = x mod 7, where x is the integer value represented by a block of 4 bits. Comment on the correctness of the execution.

Question #2 [rank 6+3+2]. Consider the **blocked** WAND algorithm (with blocks of size 3 docIDs) and assume that it is examining the heads of the following four posting lists:

t1 → (1, 5, 16, 17, 18, 21), with $ub_1 = 1$, and local upper bounds of the two local blocks = 0.5 and 1
t2 → (4, 5, 15), with $ub_2 = 0.2$, and local upper bound of the unique local block = 0.2
t3 → (5, 6, 8), with $ub_3 = 0.3$, and local upper bound of the unique local block = 0.3
t4 → (2, 5, 6, 7, 8, 9), with $ub_4 = 2$, and local upper bounds of the two local blocks = 1 and 2

At that time the current threshold equals 3.3. Answer the following questions:

1) Describe the working of the blocked WAND, show which is the next docID chosen as pivot, comment **whether or not** its full score is computed.
2) Indicate which is the docIDs examined in every list after Question 1 (*hint*: recall that one entire block is discarded…)
3) Comment how you could derive that the second blocks of t1 and t4 have local upper bounds equal to 1 and 2, respectively, even if the exercise would not have specified them.

Question #3 [rank 5]. You are given the sets A = {1, 2, 3} and B = {2, 3, 5}, with the following four hash functions $h_j(x) = j * x \bmod 7$, with j = 1,2,3,4. Estimate the Jaccard Similarity between A and B using the min-hashing approach.

Question #4 [rank 3]. Provide a graph in which the Personalized PageRank centered in the node A and computed for a node B is different of the Personalized PageRank centered in the node B and computed for the node A. Justify the answer.