Information Retrieval – exercises
8 February 2023 – time 60 minutes

Name and Surname: #matricola:

Question #1 [scores 5] Show which (triples of) items are compared by the algorithm that computes the intersection among the following three posting lists:

T1 -> 3, 5, 6
T2 -> 2, 3, 6
T3 -> 2, 3, 4, 6

Question #2 [rank 5+2]. Given the following 3 documents D1 = "a b c", D2 = "b d b", D3 = "a d a", and assuming that shingles are composed by one single letter (hit: do not consider spaces, so the first document is formed by three shingles: a, b, c)

- compute the MIN-hash sketch for the three documents above by considering 2 hashing functions h1(x) = 3*x mod 5, and h2(x) = 2*x mod 5, where x=1,2,3,4 for the encoding of a,b,c,d respectively.
- Show which is the most similar pair of documents according to the sketches above.

Question #3 [rank 4]. Show how it is compressed by the algorithm WebGraph the posting list of the node 16, with respect to the previous posting list:

15 -> 3, 5, 6, 7, 8, 10, 16, 17, 18, 22, 24, 26, 34
16 -> 5, 6, 7, 8, 9, 16, 17, 20, 21, 22, 24, 29, 30

Question #4 [rank 4+2]. Consider the Blocked-WAND algorithm for examining the head of the following four posting lists:

t1 → 5, 6, 12, 13
t2 → 2, 6, 7, 8, 11
t3 → 1, 6, 9, 13, 15
t4 → 6, 7, 8, 11

The current threshold is 2.2, and the **upper bounds** of the scores in each posting list are:
ub_1 = 0.4, ub_2 = 1, ub_3 = 0.6, and ub_4 = 0.5.
Moreover the blocks have size 3, and the **local upper bounds** of the first block of each list is
lb_1 = 0.4, lb_2 = 0.5, lb_3 = 0.3, and lb_4 = 0.5.

- Which is the candidate docID, and its full score is computed? (Motivate your answer)
- Which block is discarded to go to the next docID?

Information Retrieval – theory
8 February 2023 – time 45 minutes

Name and Surname: #matricola:

Question #1 [scores 3] Describe the two approaches to distributed indexing: Term-based vs Doc-based partitioning and highlight their computational differences in solving the queries.

Question #2 [rank 2+2+1] Define what is a Bloom Filter, which operations it supports, and its probability of false-positive error.

Question #3 [rank 2+2] Given a random walk over a weighted graph,

- state the two properties that make it converging to a unique steady state probability distribution independently by the starting distribution.
- provide two examples of graphs for which each one of these properties does not hold.