Information Retrieval – exercises
24 July 2023 – time 60 minutes

Name and Surname: #matricola:

Question #1 [scores 5+2] Consider the Blocked-WAND algorithm for examining the head of the following four posting lists:

t1 → 4, 6, 12, 13, 17, 20
t2 → 3, 6, 7, 8, 11, 15
t3 → 1, 6, 9, 13, 15, 16
t4 → 6, 7, 8, 11, 12, 13

The current threshold is 2.2, and the **upper bounds** for 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.

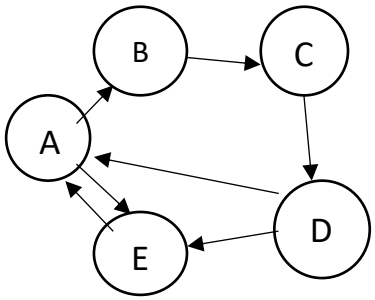
- Which is the candidate docID ? Comment the answer and state whether its full score is computed.
- Which block is discarded to go to the next docID?

Question #2 [scores 5] Given the following 3 posting lists, use the algorithm WebGraph to compress **only** the posting list of node 16, and comment the choice of which previous posting list between 14's or 15's is chosen to compress 16's one:

14 -> 3, 5, 6, 7, 8, 10, 16, 17, 18, 22, 24, 34
15 -> 2, 3, 10, 12, 14, 18, 22, 24, 25, 27, 28, 34
16 -> 5, 6, 7, 8, 9, 10, 16, 17, 20, 21, 22, 24, 30

Question #3 [scores 3] Decompress the following binary sequence 000100000110 which is known to be compressed by γ-code (i.e., gamma code).

Question #4 [scores 5] Given the following graph:



Rank the nodes according to their "similarity" to node D by means of Personalized PageRank, with uniform starting distribution and alpha = 0.5.

Information Retrieval – theory
24 July 2023 – time 60 minutes

Name and Surname: #matricola:

Question #1 [scores 3] Comment on how the TF-IDF score is stored together with the posting lists.

Question #2 [scores 3+3]

- Describe the LSH-sketch that identifies the binary vectors being similar according to the "Hamming distance". (*hint*: not include the graph/clustering; just the LSH-based approach to "match")
- State and prove the probability that two binary vectors of size d and hamming-similarity *s* are discovered to "match" by the proposed LSH-sketch.

Question #3 [scores 3] Comment on the Latent Semantic Indexing (LSI) applied to the term-document matrix A (of size m *x* n), why it has been introduced and how it is used on a query vector q.