Name and Surname: #matricola:

Question #1 [scores 4] Show how it is compressed by the algorithm WebGraph the posting list of the node 16, with respect to the one of node 15:

15 -> 1, 3, 5, 6, 7, 8, 10, 16, 17, 22, 24, 44
16 -> 2, 3, 5, 6, 7, 8, 9, 10, 16, 17, 20, 21, 22, 24

Question #2 [scores 3+4] You are given three sets A = \{2, 5, 6, 9\}, B = \{1, 2, 4\} and C = \{1, 5, 6, 9\}.
- Compute the Jaccard similarity between all pairs of them
- Approximate the Jaccard similarity via Min-Hashing, by using the following three permutations: \( \pi_1(x) = 3 \times x \mod 11 \), \( \pi_2(x) = x+5 \mod 11 \), \( \pi_3(x) = 4 \times x \mod 11 \)

Question #3 [scores 4] Given the dictionary of strings \( D = \{ \text{AAB, ABA, ACA} \} \) construct a bigram index (hence \( k=2 \)). Then given the string \( Q = \text{"BAAB"} \) use the overlap distance to filter a set of strings from \( D \) that are potential candidate for an edit distance \( e=1 \).

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

\[ t_1 \rightarrow 7, 9, 10, 11, 14 \]
\[ t_2 \rightarrow 3, 4, 6, 7, 8, 10, 11, 14, 16, 19 \]
\[ t_3 \rightarrow 6, 7, 8, 10, 15 \]
\[ t_4 \rightarrow 1, 3, 6, 8, 9, 11, 13, 14, 15, 16 \]

The current threshold is \( \theta = 2.8 \), the upper bounds of the scores in each posting list are: \( \text{ub}_1 = 2 \), \( \text{ub}_2 = 1.5 \), \( \text{ub}_3 = 0.5 \), \( \text{ub}_4 = 1 \), the blocks are of size \( 5 \), and the local upper bounds of the first block in each list are equal to \( \text{lb}_1 = 2 \), \( \text{lb}_2 = 1 \), \( \text{lb}_3 = 0.5 \), \( \text{lb}_4 = 1 \).

a) Which is the candidate docID, and is its full score computed?
b) Show the docID pointed by each iterator at the end of the Blocked-WAND step (that is, just before determining the next candidate docID).

Question #5 [scores 3] Compute one step of PageRank on the following graph by assuming \( \alpha = \frac{1}{2} \) and the starting probability distribution \( r(A)=2/4 \), \( r(B)=1/4 \), \( r(C)=0 \), \( r(D)=1/4 \).
Question #1 [scores 4] Describe the two approaches to dynamic indexing: i.e., 2 indexes and a cascade of indexes; and comment on the time complexity of inserting one document, by assuming that the collection consists of \( N \) documents (all of the same size, for simplicity) and the machine consists of an internal memory of size \( M \).

Question #2 [scores 2] What are the minimum and maximum number of integers that Simple9 can encode in a single 32-bit word, and why?

Question #3 [scores 2] Describe the champion lists approach for approximate top-K retrieval.