

Information Retrieval – EXERCISES

7 February 2024 – time 60 minutes

Name and Surname:

#matricola:

Question #1 [scores 3] Show how Consistent Hashing assigns eight items, whose IDs are {2, 4, 5, 9, 3, 8, 1, 11}, to three servers, whose IDs are {1, 2, 3}, using the hash function $h(x) = 2*x + 5 \bmod 13$.

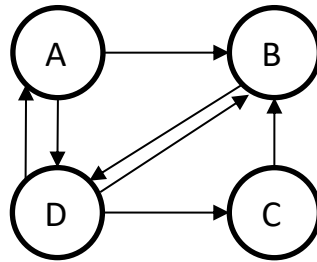
Question #2 [scores 2+2+2+2] Given a set of strings $S = \{\text{bag, bar, bus, bet, bit}\}$.

- Build a 2-gram index on S .
- Show how the 1-error search for $P = \text{"bas"}$ is executed, and specify the candidate strings.
- Select one of the candidate strings and compute the real edit distance with the pattern P by using dynamic programming.
- Can you use the Permuterm index to solve the query (b) above, and how? Motivate the answer.

Question #3 [scores 2+2+2+2] Given the sequence of integers $S = (1, 5, 7, 10, 12, 18, 21)$ show how to compress:

- S via the Elias-Fano code.
- the gap-encoded S via the gamma code.
- the gap-encoded S via the PForDelta code with base = 1 and $b = 2$.
- the gap-encoded S via t-nibble with $t = 3$.

Question #4 [scores 3] Compute the authority score and the hub score of nodes B and D in the following graph via one step of the HITS algorithm. Assume that the starting vectors of authority and hub scores are both equal to [1, 2, 1, 1].



Information Retrieval – THEORY
7 February 2024 – time 45 minutes

Name and Surname:

#matricola:

Question #1 [scores 2+2] What is a strongly connected component in a directed graph? What does it mean that the Web is a Bow Tie?

Question #2 [scores 2] Write the TF-IDF formula. When is it maximized, and when is it minimized?

Question #3 [scores 2] Let A be the binary term-document incidence matrix. What does an entry $T[i,j]$ of $T=AA^T$ represent?