Information Retrieval

5 February 2021 – time 45 minutes

Question #1 [rank 4+3]. Given the list of items S = (1, 1, 3, 2, 2, 4, 2, 1, 3), and two hash functions h1(x) = 2x mod 7 and h2 = 3x mod 7,

- Build a Spectral Bloom Filter on S, using h1 and h2.
- Comment on how to solve Query(2), and whether the answer is correct.

Question #2 [rank 5]. Given the following three adjacency lists, compress them via the Web graph algorithm by choosing always the best previous list to differentially encode the current one. ("Best" = the one that induces the most copies.)

10 → 3, 10, 11, 13, 15, 17
11 → 5, 10, 11, 13, 14
12 → 3, 10, 11, 13, 21, 25, 30

Question #3 [rank 5+5]. Given the following binary strings: A = 01000, A' = 01001, B = 01010, B' = 01101, C = 10, C' = 11.

- Construct a two-level indexing solution in which each disk page contains two strings, and it is compressed via Front Coding, and the strings indexed in internal memory are in a Patricia Trie.
- Show how it is searched the string S = 01011.

Question #4 [rank 6]. You are given the two files:

F_old = "abxxab", and F_new = "x_abab",

Assume a block size B=2 chars, and hash function $h(c_1\ c_2) = (c_1 + c_2) \bmod 7$, where we assume that {a, b, _, x} map to the values {1, 2, 3, 4}. Show the execution of the algorithm *zsync* based on that hash function.

Question #5 [rank 2]. Let us assume that you are given two terms t1 and t2, whose posting lists have length n1 and n2, respectively.

- **Write the pseudocode** to solve "t1 and (not t2)". Use NEXT[] as operator to advance on a posing list, and use HEAD[] as operator to return the head of a posting list.
- **Students of Informatica-Umanistica** can use a wordy description of the algorithm.