Information Retrieval

Midterm - 9 November 2021 – time 60 minutes

Question #1 [rank 5]. Given the two sorted lists (shown horizontally):

1	2	5	**6**	**7**	**9**	10	12	14
4	11	24						

Specify the pairs of items which are compared by the algorithm that intersects the two sorted lists (one of 9 items and the other with 3 items), by assuming that it uses **Skip Pointers** on blocks of 3 items each (shown with different background in the picture). Each last block has skip pointer equal to 50.

Question #2 [rank 6]. Consider the following adjacency lists compressed via the WebGraph compression algorithm, and decompress the adjacency lists of nodes 4 and 5.

Node	Outd	Ref	Num Blocks	Copy List	Extra nodes
3	9	0	0	-	2, 9, 10, 12, 13, 17, 19, 22, 25
4	8	1	5	0,0,1,0,3	5, 16
5	5	1	4	1, 2, 1, 0	50

Question #3 [rank 6]. Given the sets A = { 1, 5, 7, 8 } and B = { 4, 5 } and C = {5, 8, 9}, simulate the min-hashing technique in approximating the Jaccard similarity among them, by using the following three permutations $P_1(x) = x \bmod 11$; $P_2(x) = 2*x \bmod 11$; $P_3(x) = 3*x \bmod 11$.

Question #4 [rank 4+3+3+3]. Given the dictionary D = { bass, bot, box, call, cat, cow }

- Build a 2-gram index over D.
- Show the candidate strings having error e=2 with the query “bata” when using the 2-gram index of the previous point and ONLY the overlap distance.
- Show the candidate strings having error e=1 with the query “bata” when using the 2-gram index of the previous point and ONLY the overlap distance.
- Show the 2-level index (i.e., compacted Trie + Front coding) built over the strings of D by assuming a disk page able to store 3 strings.