Information Retrieval – EXERCISES
27 May 2024 – time 60 minutes

Name and Surname: #matricola:

Question #1 [scores 5] Let us be given the following adjacency list:

15 → 3, 5, 6, 7, 8, 10, 16, 17, 18, 22, 24, 26, 34

and assume that the list of node 16 is compressed by WebGraph as follows:

Outd = 13
Ref = 1
#blocks = 7
Copy blocks = 0, 0, 3, 0, 1, 0, 1 [the last value has been dropped and thus not shown]
Extra nodes = 9, 20, 21, 29, 30

derive the original adjacency list of node 16.

Question #2 [scores 5]. Given the four texts:

T_1 = "a beautiful dog"
T_2 = "after after dog"
T_3 = "after a girl"
T_4 = "girl beautiful girl"

by using the TF-IDF vectors of the four texts above (logs are in base two), find the most similar text to the query "beautiful girl" (do not apply any normalization to the cosine score).

Question #3 [scores 1+4] Suppose we wish to design a code that follows the design of Simple9 but over 2-byte words and using 3 bits for the layout part (thus enabling up to 8 configurations).

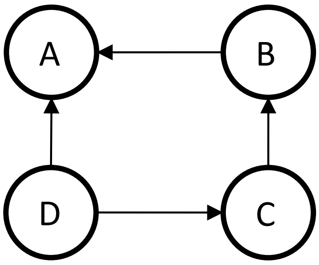
a) What is the minimum and maximum number of integers that *two* 2-byte words can encode with such a coding scheme and why?
b) Design such a coding scheme by showing its configurations, namely, by detailing the number of integers each configuration encodes and their bit-length.

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

t_1 → 5, 7, 12, 13, 15, 17
t_2 → 1, 7, 9, 13, 15, 16
t_3 → 2, 6, 7, 8, 11, 16, 18
t_4 → 7, 8, 9, 11, 13, 15

The current threshold is θ = 0.75. The upper bounds of the scores in each posting list are $ub_1 = 0.2$, $ub_2 = 0.3$, $ub_3 = 0.2$, $ub_4 = 0.2$. The blocks are of size 3, and the local upper bounds of the first block of each list are $lb_1 = 0.2$, $lb_2 = 0.25$, $lb_3 = 0.15$, $lb_4 = 0.18$. Which is the candidate docID, and is its full score computed?

Question #5 [scores 4] Compute one step of PageRank on the following graph by assuming α = 1/3 and the uniform starting probability distribution.



Information Retrieval – THEORY
27 May 2024 – time 45 minutes

Name and Surname: #matricola:

Question #1 [scores 2] Describe the soft-AND query, and comment on its design.

Question #2 [scores 3] Describe and motivate the use of the heap in Mercator.

Question #3 [scores 3] Let us fix a list T of possible topics of web pages. Suppose each user u of our search engine is associated to a vector V_u of length |T|, whose *i*-th entry is a real number between 0 and 1 representing the user interest in topic *i*, and that the entries of V_u sum to 1. For example:

T = [science, politics, entertainment]
V_{u1} = [0.3, 0.5, 0.2]
V_{u2} = [0.8, 0.1, 0.1]

Describe how to personalize the PageRank according to the interests of each user. N.B.: This must be done efficiently, thus without running a separate PageRank computation for each user.