# Information Retrieval

## Final term – 14 december 2021 – time 45 minutes

**Question #1 [rank 3+3+3].** Given the following four texts:

T1 = “white Xmas”
T2 = “Xmas Xmas happy”
T3 = “happy white”
T4 = “red red”

- Compute the Inverted List, by encoding its postings with gamma-code (no gap coding between the docIDs)
- Compute the TD-IDF vectors (log is in base 2)
- Compute and show the most similar text to Q = “red Xmas” (no norms)

**Question #2 [rank 6].** Consider the WAND algorithm by assuming that at some step it is examining the heads of the following four posting lists:

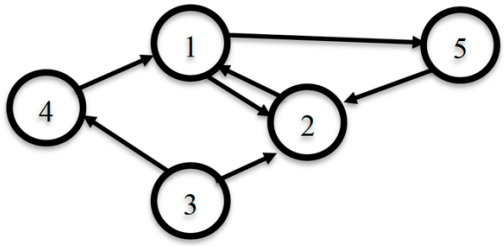
t1 → (…, 1, 5, 6, 7, 8, 11)
t2 → (…, 4, 13, 15)
t3 → (…, 5, 6, 8, 9)
t4 → (…, 2, 3, 5, 6, 7, 8, 11)

At that time the current threshold equals 3.7, and the upper bounds of the scores in each posting list are: $ub_1 = 1$, $ub_2 = 0.2$, $ub_3 = 0.6$, $ub_4 = 2$.
Which is the next docID chosen as pivot and comment **<u>whether or not</u>** its full score is computed. *(Motivate your answer)*

**Question #3 [rank 4+4].**

- Given the following three texts: T1 = “aabb”, T2 = “aaaa”, T3 = “bbbb”. Show what is the chosen compressed sequence by the algorithm z-delta, by assuming that the “cost” of a compression is estimated via the number of LZ-phrases.
- Given f_new = AAB AAB AAA BBB and f_old = AAAB, show the steps executed by zsync with block size 3 characters (*spaces in f_new are present only to ease the reading*).

**Question #4 [rank 4+3].**




- Compute the personalized PageRank for the node 5 by assuming a starting distribution [1/5, 0.0, 2/5, 1/5, 1/5] and alpha=0.5.
- Comment whether a random walk computed over this graph is converging to a single state which is independent of the starting distribution.