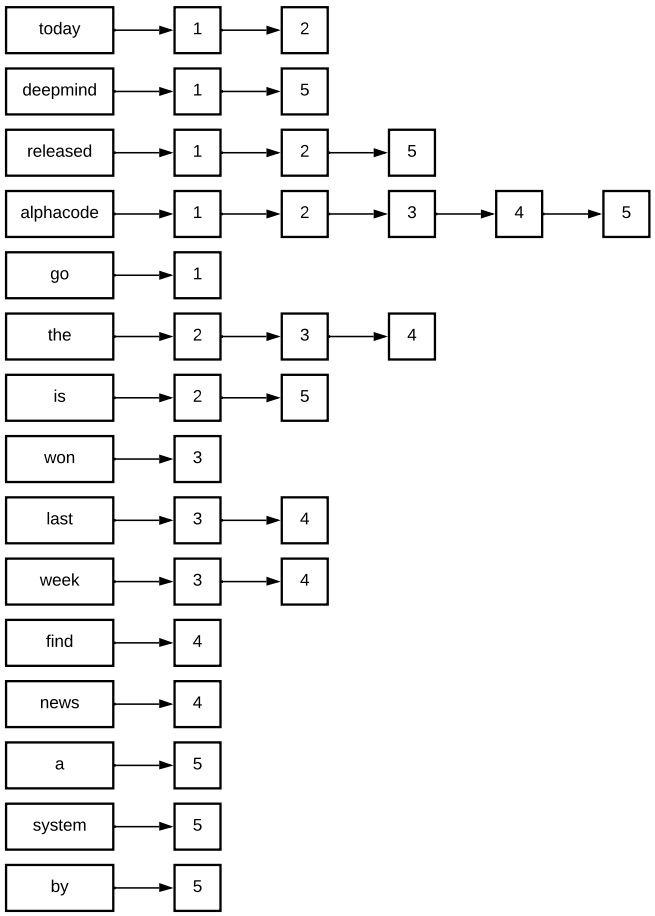
COMP 631: Homework 1 Solution

Q2

1)



Note: 1) Capital letters must be turned to lowercase ones. 2) It is fine to use the preprocessed terms 3) Initializing the document index from 0 or 1 are both fine.

2)

release: NaN

alphacode: $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$

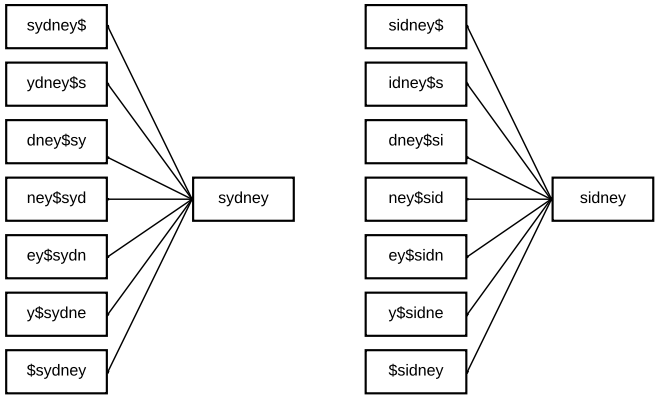
We perform "AND" operation. So no document will be retrieved.

Note: 1) "release" and "released" are different. 2) It is fine to use the preprocessed terms.

3) The inverted index is more efficient in storing large text data with sparsity.

Q3

1)



2)

s*dney → dney$s*

Then retrieve all the words with a B-tree.

3)

- "AlphaCode" and "AlphaGo" : 3 (C → G, add d, add e)
- "november" and "december" : 3 (n → d, o → e, v → c)
- "condense" and "confidence" : 3 (add f, add i, s → c)

4) "AlphaCode" and "AlphaGo" : 3

- "AlphaCode" : "Alp", "lph", "pha", "haC", "aCo", "Cod", "ode"
- "AlphaGo" : "Alp", "lph", "pha", "haG", "aGo"

"november" and "december" : 3

- "november" : "nov", "ove", "vem", "emb", "mbe", "ber"
- "december" : "dec", "ece", "cem", "emb", "mbe", "ber"

"condense" and "confidence" : 2

- "condense" : "con", "ond", "nde", "den", "ens", "nse"
- "confidence" : "con", "onf", "nfi", "fid", "ide", "den", "enc", "nce"

Note that Jaccard coefficient (the overlap ratio) is not necessary to provide here.

Q4

1)

We should first get TP, FP, and FN from the statement,

- TP = 125
- FP = 50
- TN = We don't know from the statement
- FN = 200-125 = 75

After then, we can get precision and recall as follows,

- Precision = $\frac{125}{125+50}$
- Recall = $\frac{125}{125+75}$

2) Recall rate is more crucial when we care more about the amounts of relevant documents retrieved from the given query. For example, we usually care more about the recall rate in recommender systems. In this searching scenario, we first retrieve the relevant items, and then sort those retrieved relevant items after the retrieval stage. The relevance of the retrieval items

in the retrieval stage highly impacts the recommendation results. No matter how sorting algorithm works, if the retrieved items are irrelevant, then the recommendation results are not satisfactory at all.

3) When using accuracy to evaluate the performance, the result might not directly show the effect we want.

$$\text{Acc.} = \frac{TP(1) + TN(90)}{TP(1) + TN(90) + FP(8) + FN(1)}$$

With the very high accuracy rate, which is 91%, we only have 11% for our precision rate. There is no standard answer to the examples but make sure to illustrate the negative impact from True Negative samples.