Information Retrieval

**Spring 2018**

Assignment 2: Ranked Retrieval

Mayank Bansal

A20392482

**Query Execution by Method and Length of Query**

A good way to examine all the methods is by checking the time it takes to find the results and compare it with increasing lengths of the number of terms in the query. The number in the brackets indicates the number of words after removing stop words)

queries = [  
 "with without yemen", (2 words)  
 "is germany a real country", (3 words)  
 "can germany win the war", (4 words)  
 "what do they speak in germany", (5 words)  
 "why do british people hate the germans", (6 words)  
]

After building the index:

>>> Build Index  
TF-IDF Index built in 0.70237684249877929688 seconds  
Generating document vectors  
Document vectors built in 11.93869590759277343750 seconds  
Generating document magnitudes  
Magnitudes calculated in 8.12890481948852539062 seconds  
Champions list built in 0.06831097602844238281 seconds  
Cluster Pruning index built in 0.5126569271087646 seconds

Exact Query:

N=2 Results found in 0.00452828407287597656 seconds

N=3 Results found in 0.00849795341491699219 seconds

N=4 Results found in 0.02309203147888183594 seconds

N=5 Results found in 0.02482509613037109375 seconds

N=6 Results found in 0.01668572425842285156 seconds

Champions List:

N=2 Results found in 0.00122904777526855469 seconds

N=3 Results found in 0.00462388992309570312 seconds

N=4 Results found in 0.01502704620361328125 seconds

N=5 Results found in 0.03739285469055175781 seconds

N=6 Results found in 0.01707887649536132812 seconds

Index Elimination:

N=2 Results found in 0.00105285644531250000 seconds

N=3 Results found in 0.00094604492187500000 seconds

N=4 Results found in 0.00326204299926757812 seconds

N=5 Results found in 0.00730705261230468750 seconds

N=6 Results found in 0.00280690193176269531 seconds

Cluster Pruning

N=2 Results found in 0.00197696685791015625 seconds

N=3 Results found in 0.00188589096069335938 seconds

N=4 Results found in 0.00291371345520019531 seconds

N=5 Results found in 0.00256729125976562500 seconds

N=6 Results found in 0.00357723236083984375 seconds

As we can see, clearly N increases, Index elimination and cluster pruning do much better than exact retrieval and champions list. We might have seen a drop in N=6 because of the choice of query words. It should follow the same trend for different queries.

Since we precompute document vectors and the document magnitudes, it is easy to calculate similarity between two documents. We don’t need to calculate it for every query.

**Choosing the value of R**

Since we have to build the champions list in the indexing itself, we need to be careful of the choice of the value for R. We don’t know the value of K in the beginning therefore I chose the following method of calculating R.

**r = int(sqrt(len(S))) + 10,** where S is the number of documents

For the following document collection sizes S:

S = 10, r = 13

S = 100, r = 20

S = 500, r = 32

S = 10000, r = 120

For this collection size of 423 documents we will get at least r = 30 documents

This value of **r** becomes a problem for ***K > sqrt(S) + 10.***