Using 2 of 5 late days

David Lee

lee2173

Project 1

1.
    a. (Done)
    b. Information Retrieval

```
unk-0 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-3134.html 1 -5.27224741 galago
unk-0 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-1699.html 2 -5.32447710 galago
unk-0 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-2288.html 3 -5.45138068 galago
unk-0 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-1681.html 4 -5.51583323 galago
unk-0 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-0891.html 5 -5.60211040 galago
```

Probabilistic Model

```
unk-0 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-1474.html 1 -6.16140853 galago
unk-0 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-2681.html 2 -6.25878150 galago
unk-0 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-2560.html 3 -6.70818281 galago
unk-0 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-1935.html 4 -6.73494840 galago
unk-0 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-2836.html 5 -6.74741805 galago
```

2.
    a. 3204 documents
    b. 15004 unique words
       118.90 Average
    c. Highest Length Document:
       ID = 1667
       Name = CACM-1781
       Length = 1560
       Lowest Length Document:
       ID = 2030
       Name = CACM-1634
       Length = 14

d. TF for science = 154
DF for science = 63

e. IDF = 5.65

f. cos sim = 0.039239

g. RSV = 23.54

3.

a. Query11:

```
11 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-2699.html 1 -6.27851314 galago
11 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-2906.html 2 -7.28502872 galago
```

Query23:

```
23 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-3148.html 1 -5.66679830 galago
23 Q0 /home/u92/lee2173/cs473/project1/./corpus/CACM-2849.html 2 -5.66868529 galago
```

b. *Code is very slow to run, it is still computing if numbers keep printing.
Cosine:
MAP = Inf
NDCG = 0.47849
P@20 = 0.075
RSV:
MAP = 0.03587
NDCG = 0.17678
P@20 = 0.025

c. I think the NDCG performed better than the other models in my code. Although there shouldn't be a big difference, I saw a bigger difference in performance. The NDCG seemed to do better because I didn't have much precision near the top of the ranks, therefore the MAP was too small and the P@20 was almost nonexistent. However, I think it may be that I had some good results in the middle of the ranks, which gave me a higher NDCG value.