CA4009: Search Technologies
Laboratory Session 3
9th November 2021

Joanna Talvo - 18342523
Chloe Ward - 18302716

3.2 Basic Experimental Procedure

BM25 k=1.2 b=0.75
BM25 k=1.5 b=0
BM25 k=1.5 b=1
BM25 k=3 k= 0.10

BM25 k=1.2 b=0.75

DOCUMENT **ALL**

Output type	**Values**
runid	BM25.1.2.0.75
map	0.2153
Rprec	0.2647
P_5	0.4560
P_10	0.4120
P_15	0.3653
P_20	0.3380
P_30	0.2940
P_100	0.1748
P_200	0.1215
P_500	0.0709
P_1000	0.0438

DOCUMENT **301**

Output type	**Values**
runid	BM25.1.2.0.75
map	0.0318
Rprec	0.1266
P_5	0.2000
P_10	0.2000
P_15	0.3333
P_20	0.3000
P_30	0.2667
P_100	0.2300
P_200	0.1800
P_500	0.1240
P_1000	0.0840

DOCUMENT **302**

Output type	**Values**
runid	BM25.1.2.0.75
map	0.5292
Rprec	0.5844
P_5	0.8000
P_10	0.9000
P_15	0.8000
P_20	0.7000
P_30	0.7000
P_100	0.5200
P_200	0.3000
P_500	0.1240
P_1000	0.0620

DOCUMENT **310**

Output type	**Values**
runid	BM25.1.2.0.75
map	0.1426
Rprec	0.2308
P_5	0.2000
P_10	0.3000
P_15	0.2000
P_20	0.1500
P_30	0.1333
P_100	0.0400
P_200	0.0250
P_500	0.0100
P_1000	0.0060

BM25 k=1.5 b=0

DOCUMENT **ALL**

Output type	**Values**
runid	BM25.1.5.0.0
map	0.2004
Rprec	0.2533
P_5	0.4400
P_10	0.3940
P_15	0.3520
P_20	0.3230
P_30	0.2780
P_100	0.1648

P_200	0.1159
P_500	0.0684
P_1000	0.0419

DOCUMENT **301**

Output type	**Values**
runid	BM25.1.5.0.0
map	0.0183
Rprec	0.0928
P_5	0.2000
P_10	0.4000
P_15	0.2667
P_20	0.2500
P_30	0.1667
P_100	0.1400
P_200	0.1300
P_500	0.0920
P_1000	0.0660

DOCUMENT **302**

Output type	**Values**
runid	BM25.1.5.0.0
map	0.4473
Rprec	0.5195
P_5	0.8000
P_10	0.8000
P_15	0.7333

P_20	0.6500
P_30	0.6000
P_100	0.4700
P_200	0.2850
P_500	0.1160
P_1000	0.0580

DOCUMENT **310**

Output type	**Values**
runid	BM25.1.5.0.0
map	0.1480
Rprec	0.1953
P_5	0.2000
P_10	0.3000
P_15	0.2000
P_20	0.1500
P_30	0.1000
P_100	0.0400
P_200	0.0250
P_500	0.0100
P_1000	0.0060

BM25 k=1.5 b=1

DOCUMENT **ALL**

Output type	Values
runid	BM25.1.5.1.0
map	0.2040
Rprec	0.2505
P_5	0.3560
P_10	0.3120
P_15	0.3013
P_20	0.2780
P_30	0.2507
P_100	0.1626
P_200	0.1118
P_500	0.0689
P_1000	0.0421

DOCUMENT **301**

Output type	Values
runid	BM25.1.5.1.0
map	0.0266
Rprec	0.1160
P_5	0.4000
P_10	0.3000
P_15	0.2667
P_20	0.2500
P_30	0.2667
P_100	0.2000
P_200	0.1600
P_500	0.1180
P_1000	0.0770

DOCUMENT **302**

Output type	Values
runid	BM25.1.5.1.0
map	0.5161
Rprec	0.5455
P_5	0.6000
P_10	0.8000
P_15	0.7333
P_20	0.7000
P_30	0.7667
P_100	0.4800
P_200	0.3200
P_500	0.1280
P_1000	0.0640

DOCUMENT **310**

Output type	Values
runid	BM25.1.5.1.0
map	0.0703
Rprec	0.1538
P_5	0.2000
P_10	0.2000
P_15	0.1333
P_20	0.1000
P_30	0.0667
P_100	0.0300
P_200	0.0200
P_500	0.0080
P_1000	0.0070

BM25 k = 2 b = 0.75

DOCUMENT **ALL**

Output type	**Values**
runid	BM25.2.0.0.75
map	0.2058
Rprec	0.2496
P_5	0.4720
P_10	0.4040
P_15	0.3507
P_20	0.3300
P_30	0.2867
P_100	0.1676
P_200	0.1160
P_500	0.0689
P_1000	0.0426

DOCUMENT **301**

Output type	**Values**
runid	BM25.2.0.0.75
map	0.0244
Rprec	0.0949
P_5	0.6000
P_10	0.4000
P_15	0.2667
P_20	0.2500
P_30	0.2000
P_100	0.2100

P_200	0.1450
P_500	0.0920
P_1000	0.0920

DOCUMENT **302**

Output type	**Values**
runid	BM25.2.0.0.75
map	0.5325
Rprec	0.5210
P_5	0.8000
P_10	0.8000
P_15	0.7333
P_20	0.7500
P_30	0.6000
P_100	0.5100
P_200	0.2650
P_500	0.1160
P_1000	0.0580

DOCUMENT **310**

Output type	**Values**
runid	BM25.2.0.0.75
map	0.1543
Rprec	0.2308
P_5	0.4000
P_10	0.3000
P_15	0.2000
P_20	0.2000
P_30	0.1333

P_100	0.0400
P_200	0.0250
P_500	0.0100
P_1000	0.0060

Observation:

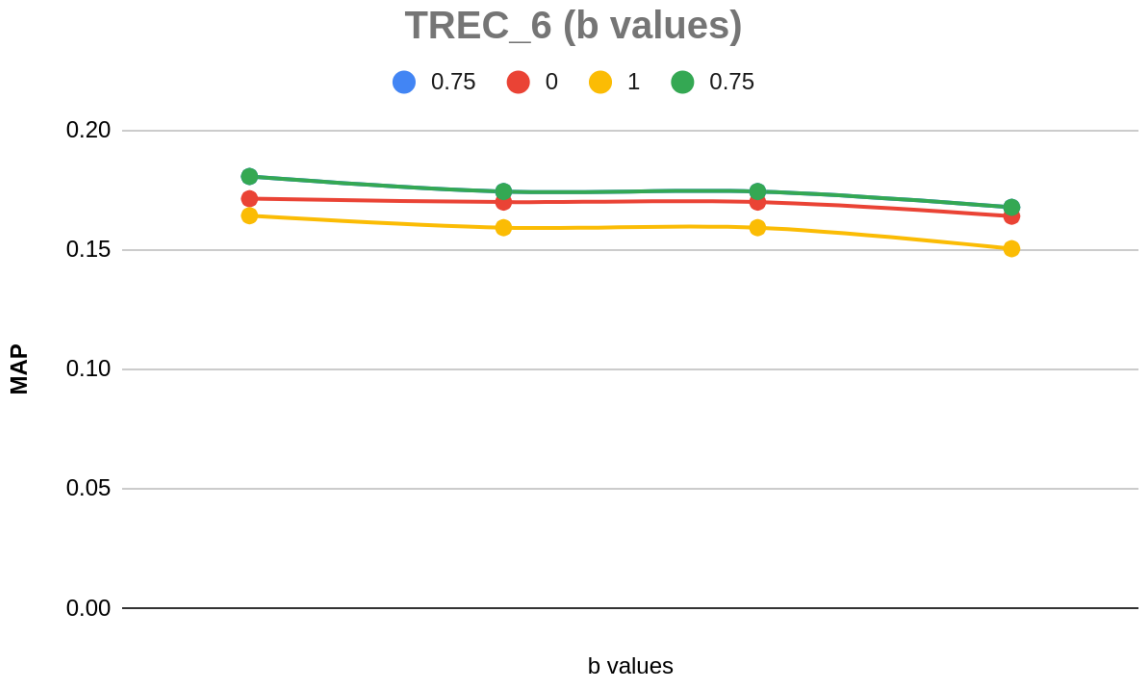
When inputting our k and b values, we see the difference between the MAP values. During our observations of the results, we realised that the MAP value which is the Mean Average Precision, when k value increases, it means that it would then search for more occurrences of the word and the results would be generated based on this MAP value.

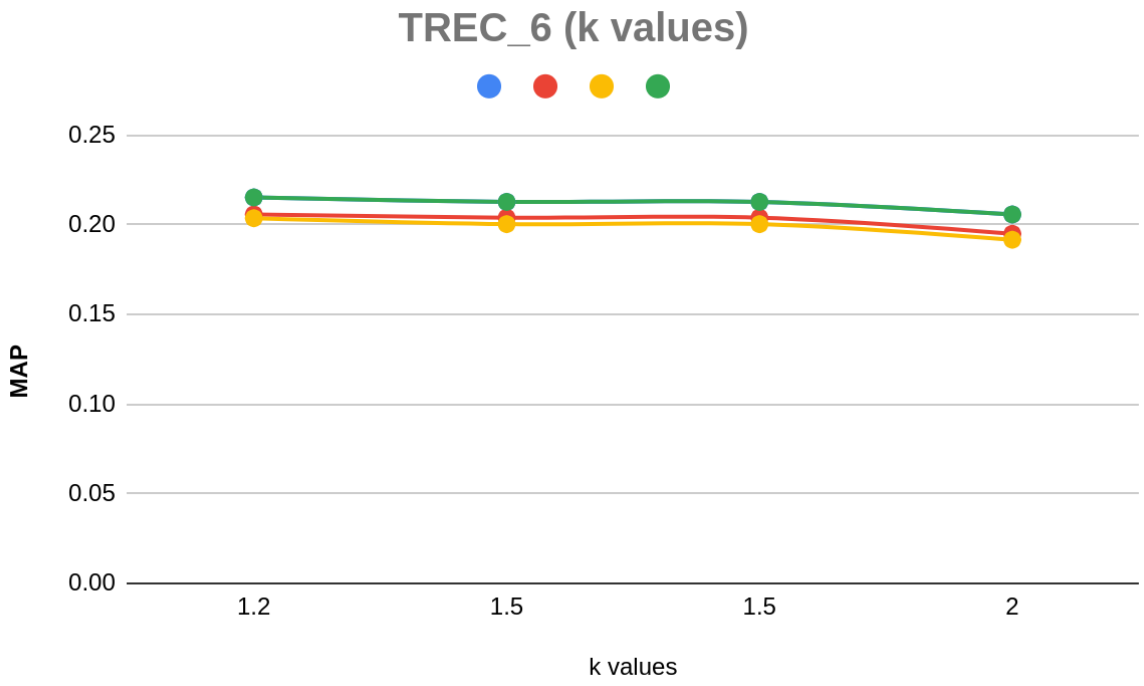
We observed that when we used k=1.2 b=0.75 and k=2 b=0.75 the MAP value would be decreased which meant that it was less accurate. This highlighted to us that the larger the K value then the less accurate the results were. This is because when you have a higher K value it will search for more occurrences of the word which will lead to the number of relevant results being impacted.

We then decided to have k=1.5 b=0 and k=1.5 b=1. We observed that when the b value was higher, the MAP values were slightly higher. This is because when you change the b value it changes the rate of normalisation for the data. When the value of b was lower the normalisation rate will be altered so it will ultimately display the results of documents that are longer, whereas if the b was higher, it will get an even amount of documents with varying lengths.

3.3 Optimisation of Parameter Values

We used the test.sh file that we were supplied with to generate the results that will be shown below.



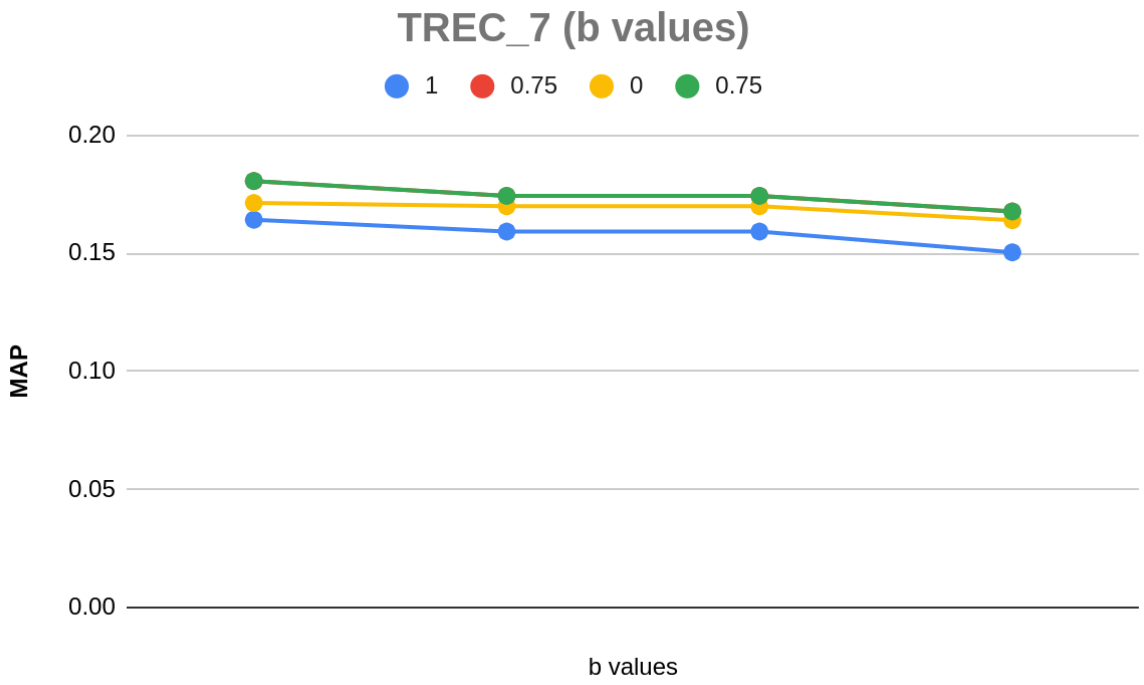


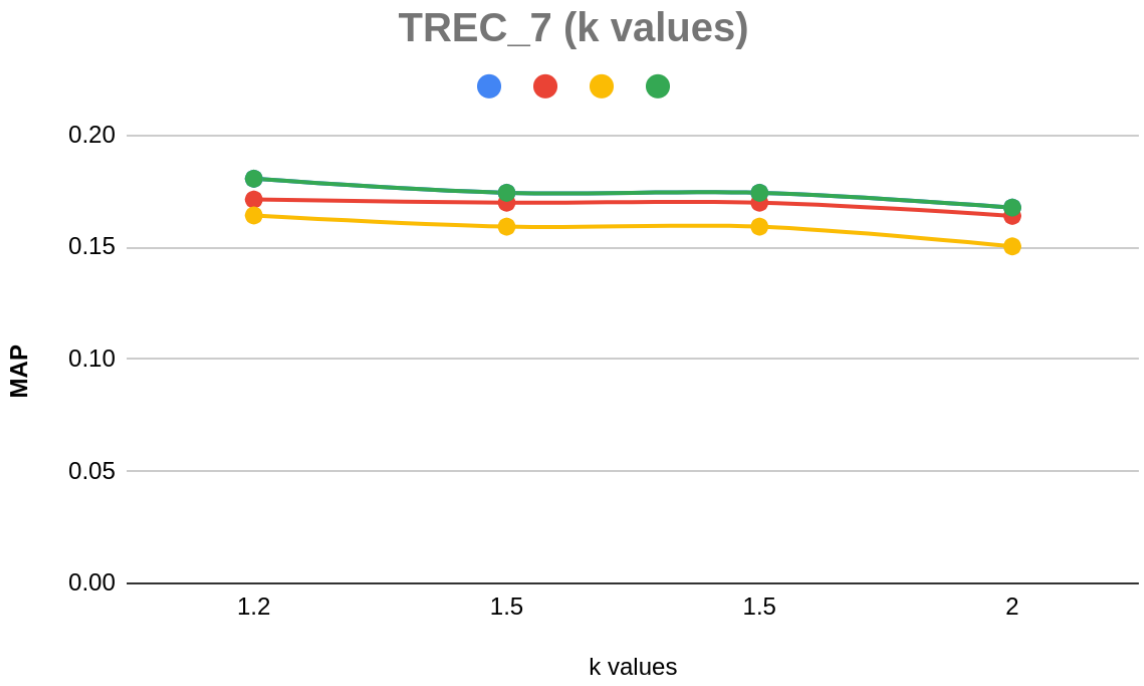
note: for b values, blue isn't visible because it overlaps with green since they have the same map values because of the same number of k and b values. And for k values, blue isn't visible because it overlaps with red.

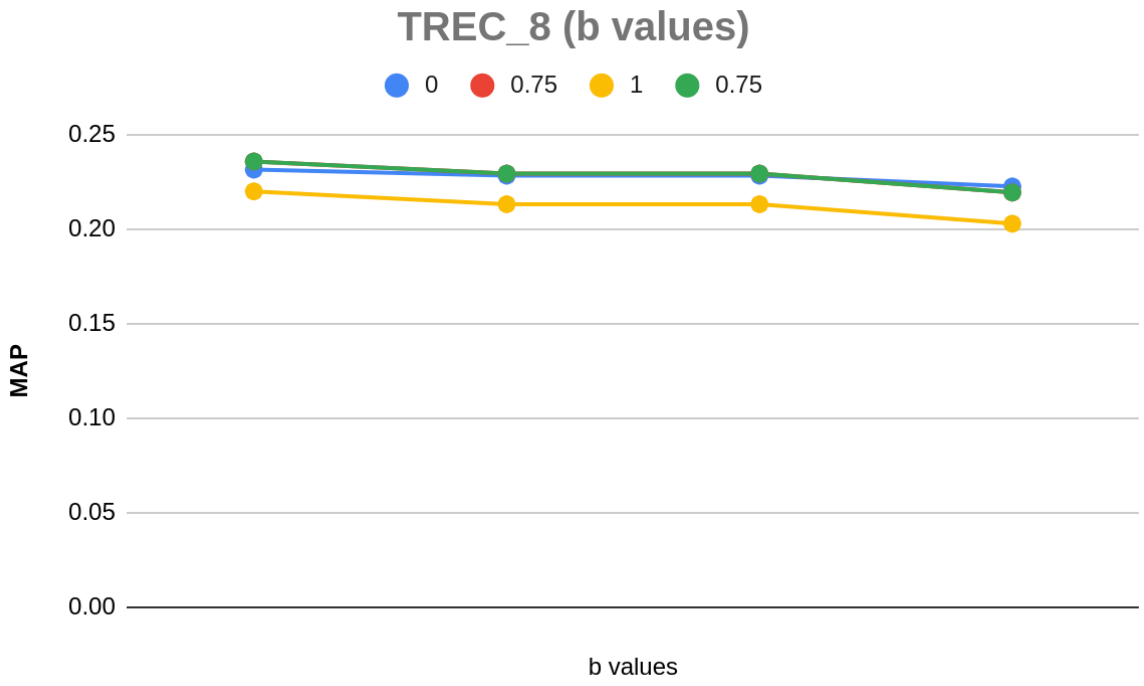
From the TREC_6 (b values), we noticed that changing the b values from 0 to 1, decreased the MAP values significantly. However with 0.75, the MAP values are significantly higher

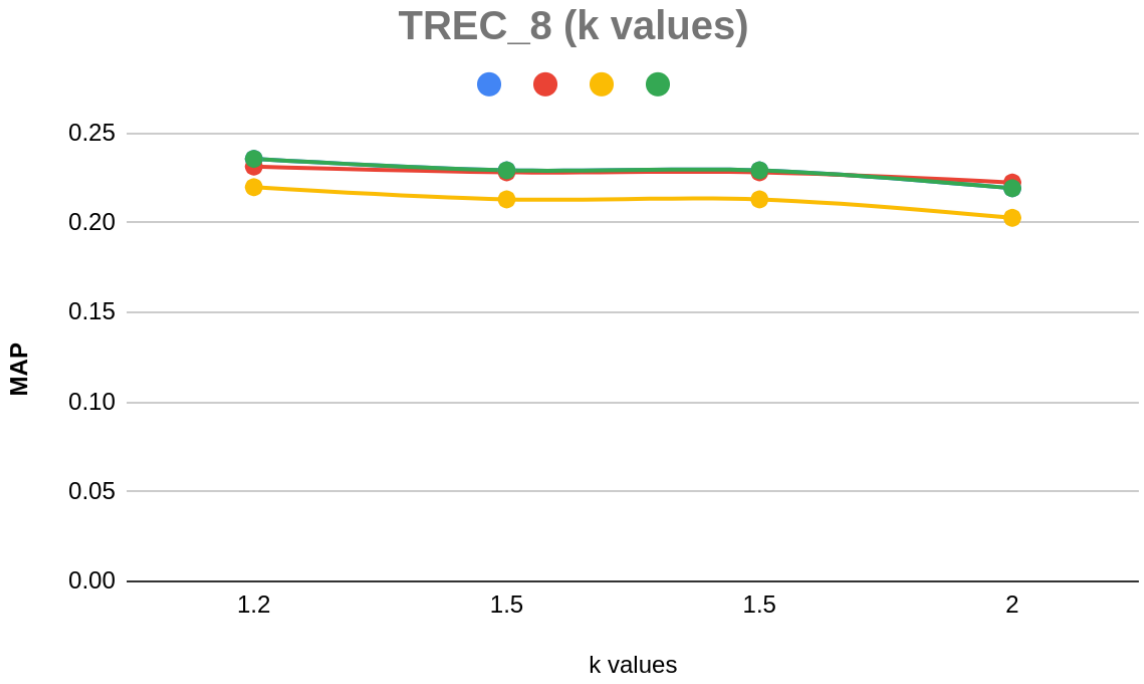
than the other two values. b values need to stay in between 0 and 1. Based on our observation, 0 and 1 as b values are not good enough for experimenting to get good results. 0.75 gives better results which we think is the reason why 0.75 is a good and standard value for b. Since this document collection is news articles, the documents have several different topics and it is often beneficial to choose a larger value for b in order for irrelevant topics from a search query to be penalized.

From the TREC_6 (k values), we noticed that 1.2 value has the higher MAP values compared to the other k values we tested which are 1.5 and 2. The higher the k value, the MAP value seems to decrease. The reason behind this is that, since we're gathering data from a collection of news articles, a short news article would very likely to have (for example our search query is "virus") the word "virus" many times with it being related to virus as a subject of the article. So for short news articles, k value should generally tend towards a smaller value in order to get a more accurate MAP.




TREC_7 (k values)
0.20
0.15
0.10
0.05
0.00
MAP
1.2
1.5
1.5
2
k values


TREC_8 (b values)
0
0.75
1
0.75
0.25
0.20
0.15
0.10
0.05
0.00
MAP
b values



There is a similarity between the three TREC values from our observation in our diagrams. It is clear that when b values are in the range 0 to 1, the MAP values are higher and when k values are smaller, the MAP values are higher. There is no drastic change in the optimal values of k and b.

The values that we would recommend for both b and k for this dataset is that considering the data that we are viewing is news articles then we would have a low value of k, such as k = 1.2 and b = 0.75, which are the standard values for BM25. We recommend these values because after looking at the diagrams, it was evident that these values of b and k gave the best MAP value which ultimately gives the most accurate results. We could have tested more values to test the different variations of the MAP values based on these parameters but for now we can definitely conclude that these standard parameter values are what work best for this type of dataset.