Optimization of K and B values in BM25 to achieve highest MAP value.

I decided to automate the procedure from the beginning using a bash script. I plotted my data using gnuplot (images attached). For my script I build a format for the URL that would need to be queried, with values such as the TREC query set, B value and K value being passed in as variables. I used a first loop to loop through each query set. On each loop, I test it with each value of K from 0.2 to 2.0 at increments of 0.2. For each of these increments of K, I then applied each value of B from 0.1 to 0.9 at increments of 0.1.

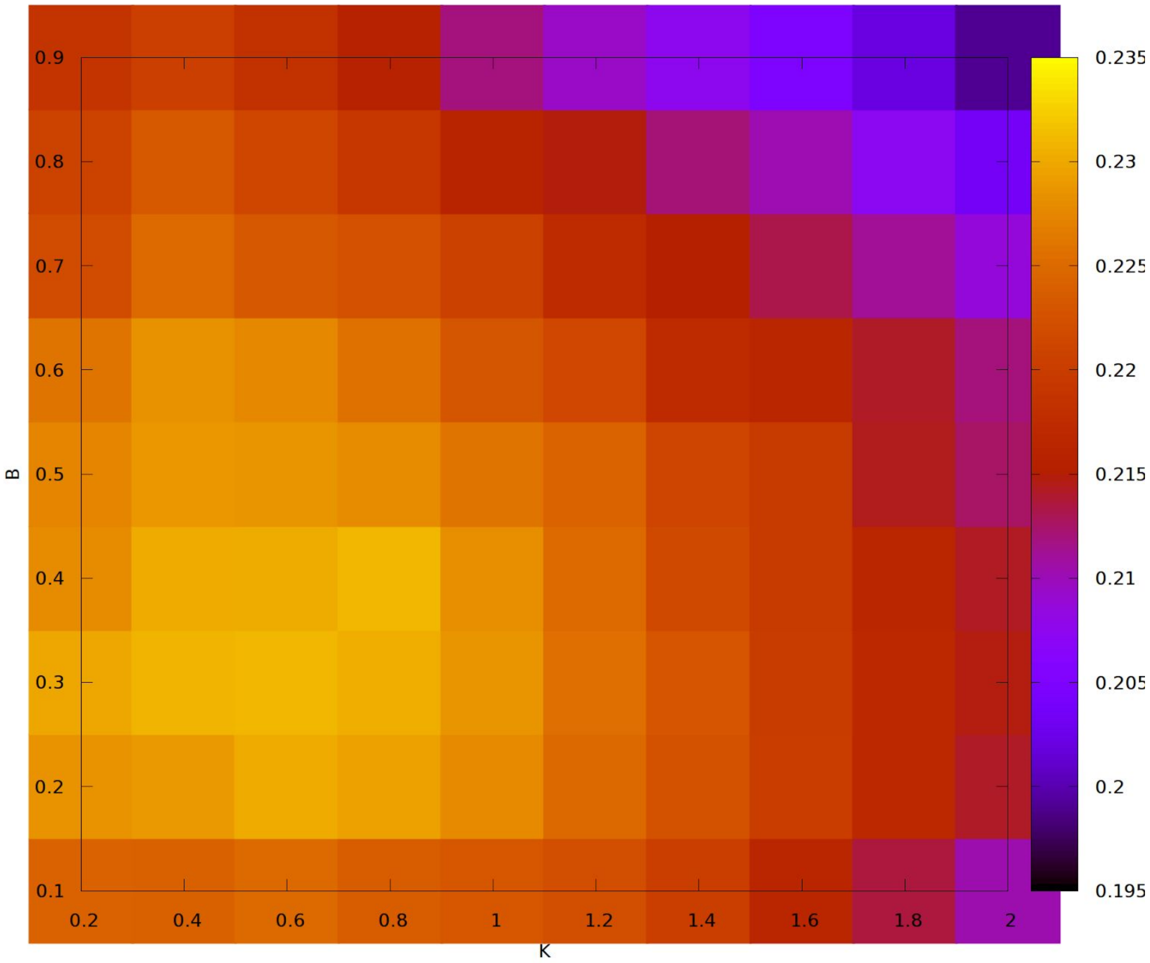
At every permutation, I called wget to get the initial reply from the server containing a link to the actual data. I then passed this through grep to find the line with the new URL ending, and parsed this with a sed to extract the URL only. This was then appended to the base URL for the website and wget was called again to get the actual output data. This output data was then grepped to get the line containing the MAP value, and sed was once again used to extract the actual value.

Once the MAP value had been retrieved, I entered the K value used, the B value, and the resulting MAP value into a new line in a file corresponding to the current TREC query set. Once all values had been retrieved for a particular set, the results were passed into gnuplot to generate a heat map of the results.

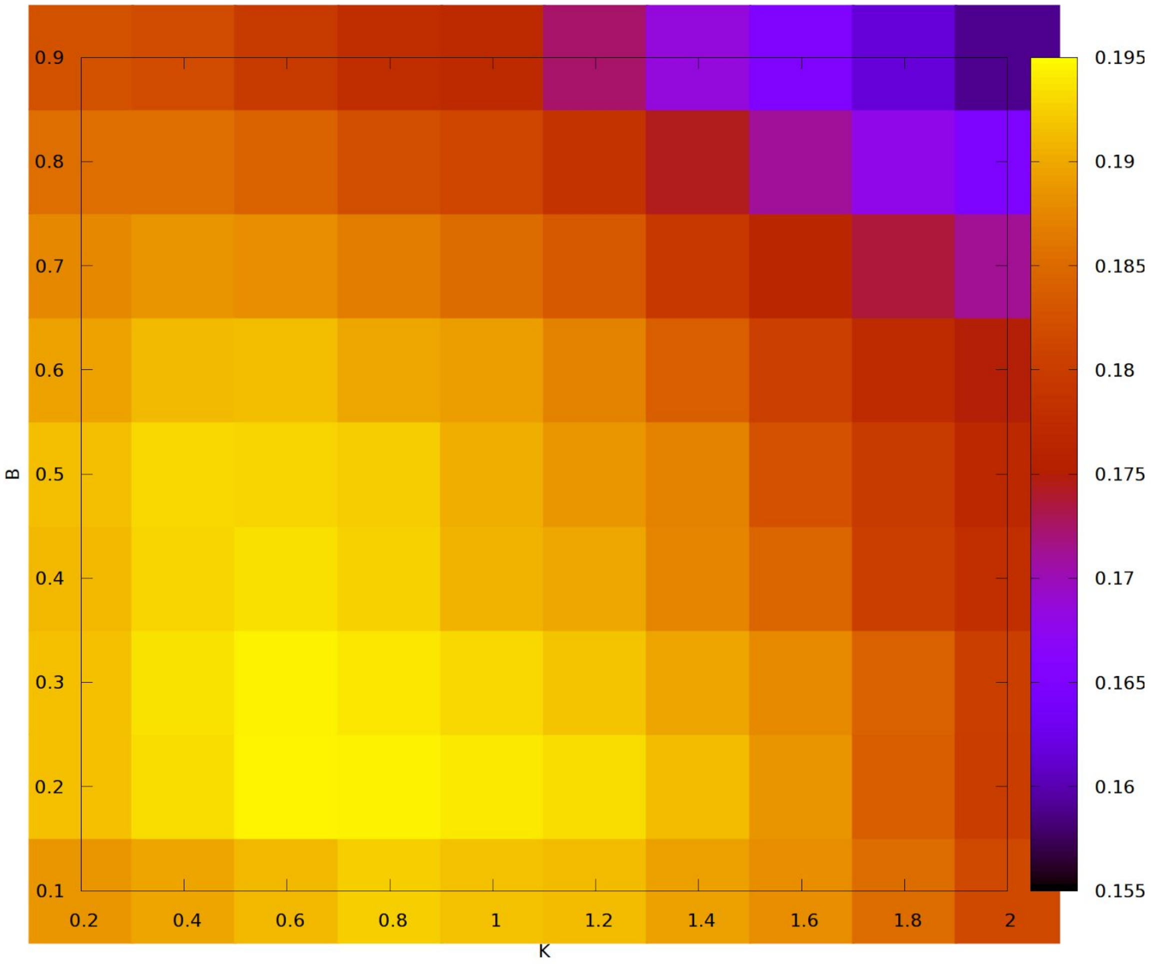
I chose to use a heat map instead of a line graph as I felt it was easier to see the distribution of results and to see any trends which occur across the variables. Code used to generate the maps is attached at the bottom of this document.

I generated a heat map for each set of the TREC queries. I then compared the results across all 3 to pick optimal values for K and B.

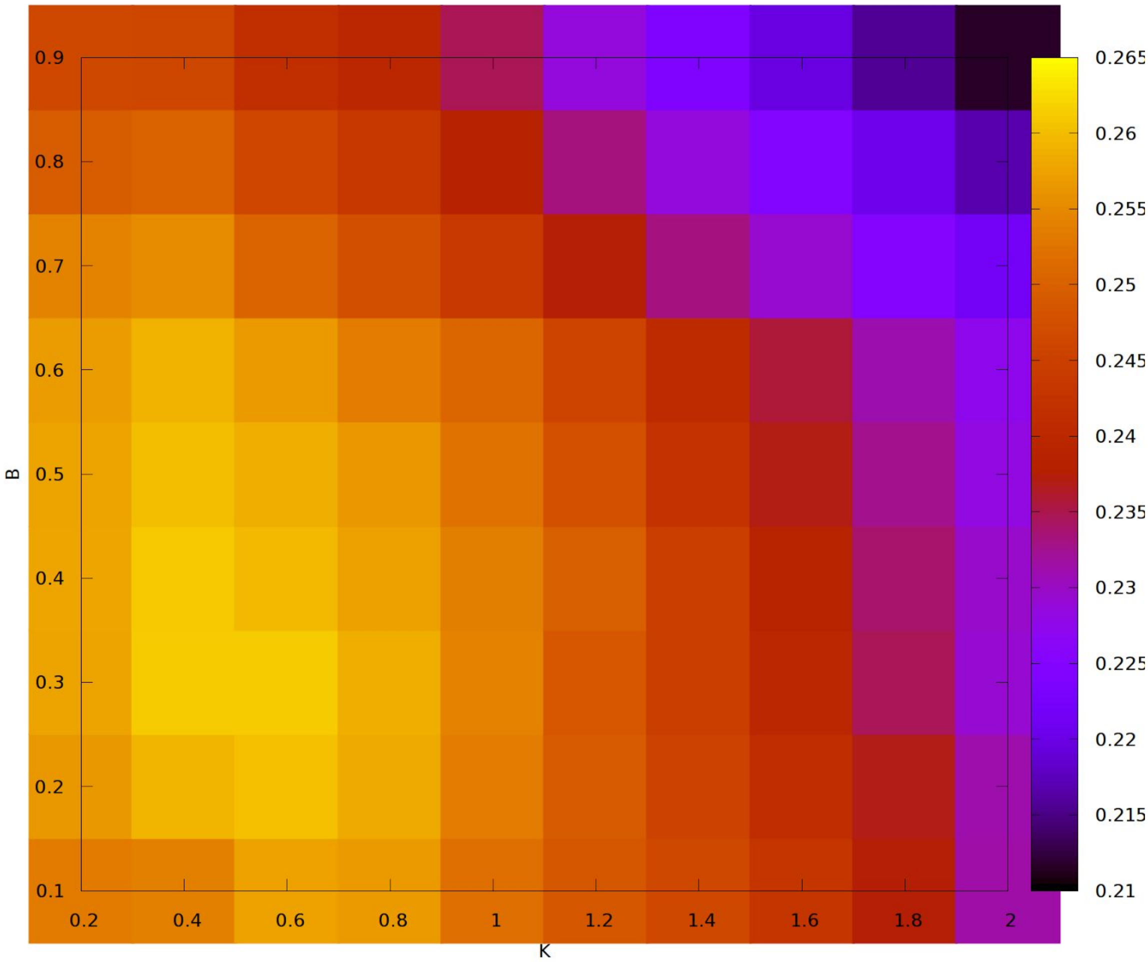
The results can be seen below.



Output from the TREC6 queries suggests (K, B) values of (0.6, 0.3).Output from the

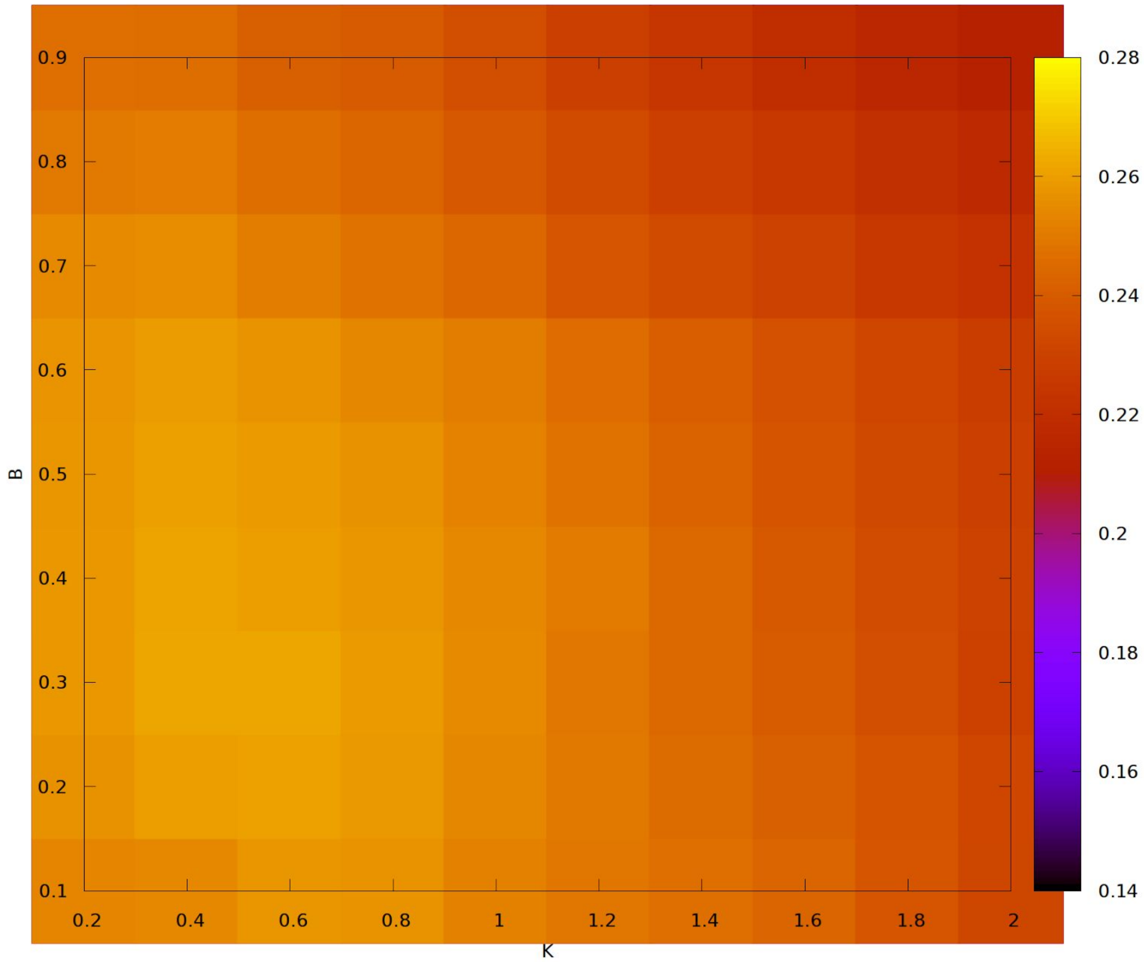

B
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
K
0.195
0.19
0.185
0.18
0.175
0.17
0.165
0.16
0.155

Output from TREC7 queries suggests (K, B) values of (0.6, 0.2)



Output from the TREC8 queries suggests (K, B) values of (0.4-0.6, 0.3).

Since there are mixed values it seems like a value 0f K=0.6 B=0.3 would be the best overall for all 3 query sets. To further test this, I concatenated the values results I obtained from all 3 sets into one file, and generated the following gnuplot heatmap:



This heatmap also seems to suggest optimal values of 0.6 and 0.3, so I would use these values in the future.

```
// Code for gnuplot
// It assumes the filename containing the data has been specified
// at the terminal before being run. This is done in the bash script
set xlabel "K"
set xrange [0:2]
set xtics 0,.2,2
set ylabel "B"
set yrange [0:1]
set ytics 0,.1,1
set zlabel "Map"
set zrange [0:1]
set ztics 0,.05,1

set autoscale
set key off

set view map
set size ratio .9

set object 1 rect from graph 0, graph 0 to graph 1, graph 1 back
set object 1 rect fc rgb "black" fillstyle solid 1.0

set terminal pngcairo size 1800,1800 enhanced font 'Verdana,20'
set output 'temp.png'

splot filename using 1:2:3 with points pointtype 5 pointsize 18 palette linewidth 50
```