CA4009: Search Technologies
Laboratory Session 3
17th November 2016

Jordan Healy - 13379226
Tríona Barrow - 11319851

Investigating BM25 Parameter Settings

- *Experimental Procedure*

BM25 k = 1.2, b = 0.75

All	0.2153
301	0.0256
308	0.4503
312	0.7433

BM25 k = 1.5, b = 0

All	0.2045
301	0.0266
308	0.4686
312	0.7573

BM25 k = 1.5, b = 1

All	0.2041
301	0.0103
308	0.3345
312	0.7433

BM25 k = 3, b = 0.75

All	0.1947
301	0.0223
308	0.3407
312	0.7433

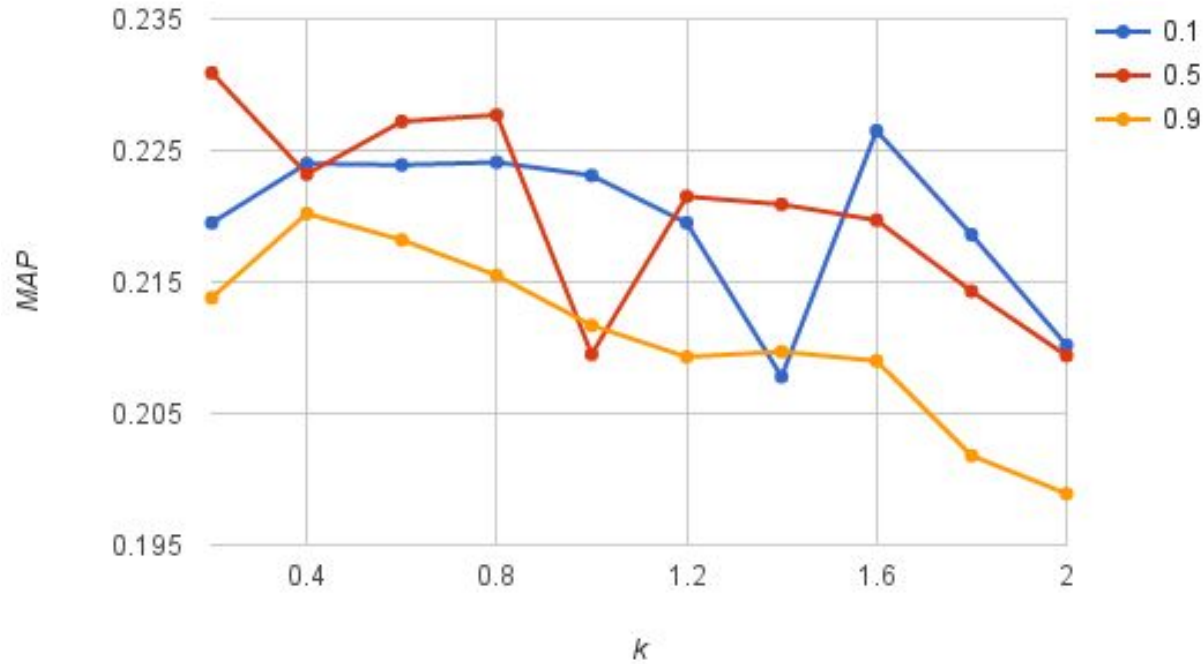
We found that the MAP (Mean Average Precision) results for the results tended to change positively within a certain range of k - when this increases then it searches for more occurrences of the words and generates the results based on this. The MAP for the overall results is affected by the number of relevant results returned as a proportion of the number of results returned - so will decrease if the results are more inaccurate. Adjusting b did not affect the results so much, as this only changes the rate of normalisation for the document length.

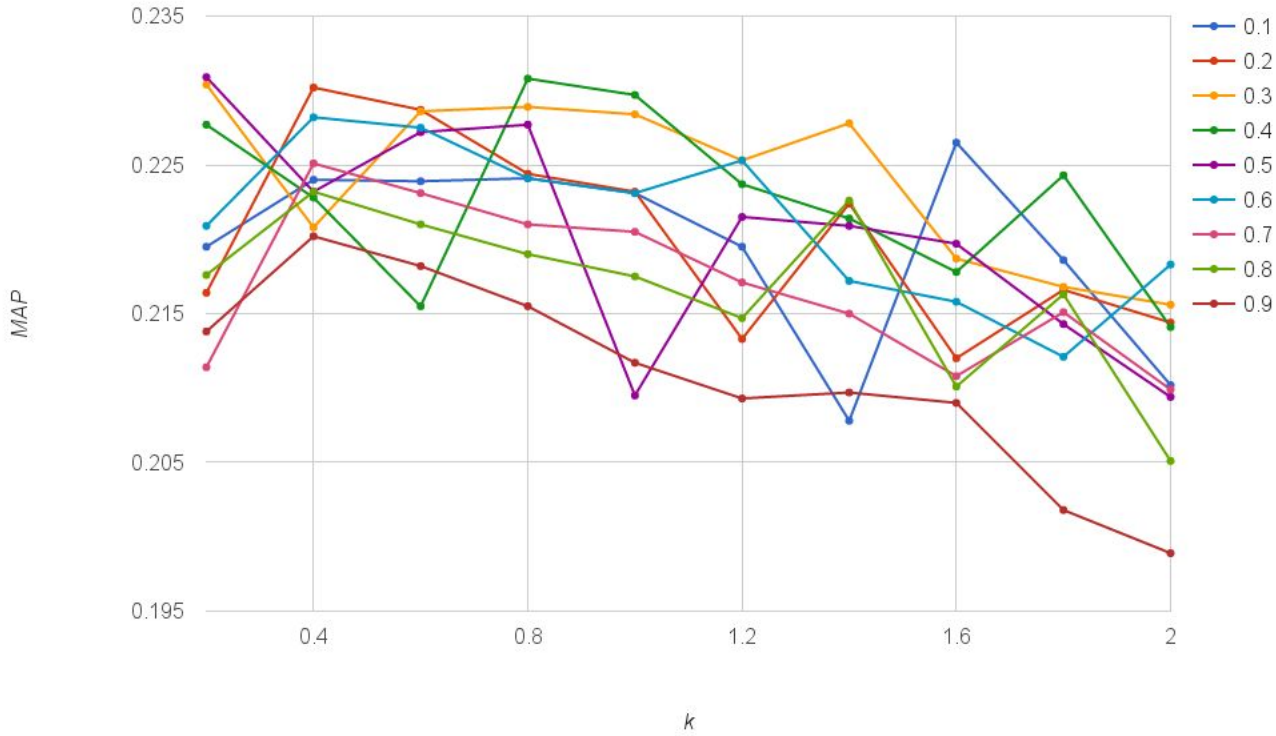
On a document basis, the MAP value suggests it's much less accurate when k is larger. As the query searches for more occurrences of each word in the query - this affects the number of relevant results returned negatively. With a smaller k value - it will search for less occurrences, so should be able to match up to the document title more accurately.

By changing the value of b, we can change the rate of normalisation for the documents retrieved. With a higher rate, normalisation attempts to generate an even value for comparing all the document lengths. With a lower value for b, the normalisation rate decreases meaning the documents retrieved with favour longer documents. Inversely, the higher the rate for b, the more evenly spread the document lengths are.

- *Optimisation of Parameter Values*

  We generated the following results using a bash script:





As we can see, we have the k values on the x -axis and their corresponding MAP values on the y-axis. Each line in our chart is a different value for b. The default value for b is 0.75 so we will compare the other values to the line “0.7”.

Initially, we see that when b=0.9 the precision values falls considerably. Also, we find it interesting that when b=0.5 and k=1.0, the precision is low, but when k is increased to 1.2 the precision value increases. This may be due to the fact that when k is 1, the results try to match exactly 1 occurrence of the query, whereas when k is 1.2, in can match to slightly more than one occurrence. The maroon line seems to be the only one that seems to follow an expected trend - with a relatively small; jump when the number of term occurrences is increased marginally, then a relatively small decrease as this is increased further.

Our recommended values for recommended values for k and b would be 0.9 and 0.4 respectively for this collection - based on the analysis of all the MAP results at each interval this seems to generate the most accurate results.