MSCI 541 – Search Engines

Homework 4

Professor Mark Smucker

Matthew Erxleben

ID: 20889980

Date: November 21st, 2023

Problem 1)

Two documents that have the same BM25 retrieval score may differ in terms of relevance. This is due to the definition of document relevance. A document is deemed relevant due to the user of the query's own discretion. This is the only way that we can define relevance, as whether the document is relevant or not is only up to the user. Therefore, even if BM25 gives two documents the same score, one user may find one document relevant but not the other.
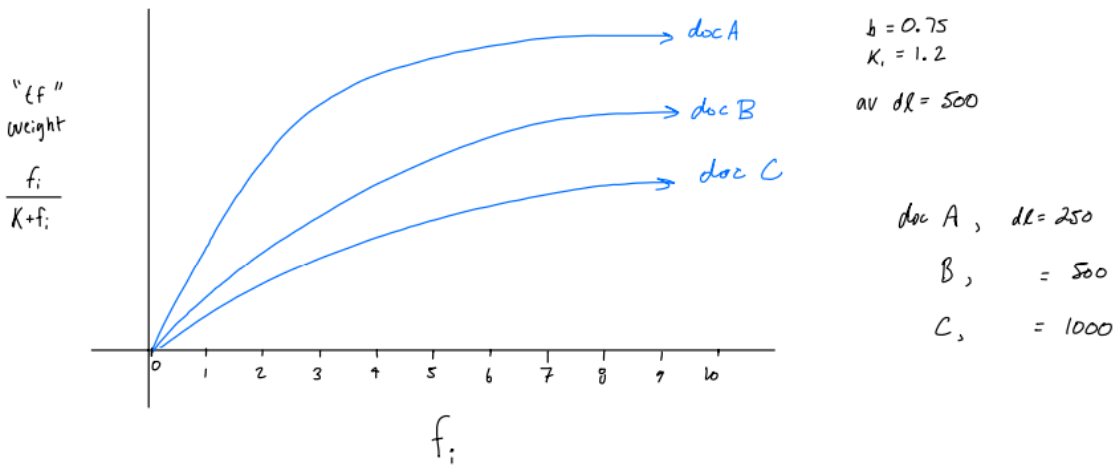
Problem 2)

2a)

In the equation for document term weights, the tf component contains a variable f_{jk} to represent the number of occurrences of term k in the document i. We use the log of this instead of just the variable in the formula. This is typically with the log base e, to be the natural logarithm. The reason for this to indicate the importance of the term k in the document i, versus just the raw number of how many times it occurs. The natural logarithm captures the exponential growth. If only the frequency was used, documents with more frequency would automatically be worth much more in the equation, which would out scale other documents with less words. The more important metric is indicating importance for a terms weight in a document, therefore that is why the log is used.

2b)

In the BM25 formula, the term frequency section uses a similar methodology. The frequency of term k in doc i is divided by k plus the frequency again. K is defined as a length normalizing coefficient, and is calculated using the doc length of doc i and the average doc lengths in the collection, as well as two tuning parameters b and k1. B is from 0 to 1, and controls the amount of length normalization (0.75 is typically used). K1 controls the saturation for term frequency in the doc (1.2 is typically used). This is similar to the previous implementation in the document term weights formula because it scales the frequency of terms appearing in the doc with how long the document is, and how long the average document is in the collection.

As we can see, the term frequency weight versus the frequency of the term is seen in the above chart, with corresponding document lengths. The term frequency weight is worth less when the document is greater in length.

Problem 3)

Typically, the frequency of terms in a query are ignored when doing BM25. However, we can use the query term frequency in the BM25 algorithm if wanted. We can do this by multiplying the terms contribution to the sum by the number of times in appears in the query. For example, if the query is "The brown Cleveland brown is brown in a brown town" and mentions the term brown four times, this would be valued differently than the query "The Cleveland brown is in a town" which mentions the term brown only once. Currently in our BM25 formulation, these queries are valued the same.

Problem 4)

The first step towards handling these words is would be stemming, if that is in the design of the algorithm. Stemming allows for words to be broken down into their basic form, and therefore increasing the chance that it may be seen in the document collection. For example, none of the documents may contain the term "eating", but after stemming this word will become "eat". This word may now be contained in a document that is in the collection; therefore, we consider that word as the query term instead. If stemming is not including in the algorithms design or still does not exist in the document collection, these words are typically handled by removing the words

from the query. This allows for certain algorithms to not get hurt from the term being included in the query terms, such as BooleanAND (which relies on all terms appearing at least once in each document)

Problem 5)

We never store this matrix as it is way to big. Instead, we use the count of terms from our inverted index to computer weights on the fly.

Problem 6)

Gary's has made many mistakes in his implementation of BM25.

The first mistake he made is converting all non-letters to spaces. This can cause words to break into words and remove the original term from the term list. For example, u.s.a will become three separate terms "u", "s", and "a" rather than becoming one term "usa". Gary would need to also include the full term "usa" for better results. Another mistake that Gary makes is not retaining the original tokens after stemming. It is important to test how the BM25 algorithm is with the original tokens, and then with the porter stemming to see what types of queries do better with the original tokens instead.

Problem 7)

Installation Requirements:

1.   Please make sure Python is installed on your computer before running the program.

2.   Clone the repository on your device by entering this into your terminal: git clone https://github.com/UWaterloo-MSCI-541/msci-541-f23-hw4-matterxleben.git

Running the Programs:

To run these programs, please navigate to where you cloned the repository and open the working directory .../msci-541-f23-hw3-matterxleben

IndexEngine:

This program accepts two command line arguments:

1.   a path to the latimes.gz file

2.   a path to a directory where the documents, metadata, and term files will be stored.

For example, you would run IndexEngine from the command prompt / terminal / shell as:

python IndexEngine.py C:/Users/matth/OneDrive/Desktop/University/3B/MSCI541-Search-Engines/HW4/raw-data/latimes.gz C:/Users/matth/OneDrive/Desktop/University/3B/MSCI541-Search-Engines/HW4/store

ResultsEvaluation.py:

The program accepts 2 command line arguments: the directory location of your results file, the directory location of your qrels file:

For example, you would run ResultsEvaluation.py from the command prompt / terminal / shell as:

    python ResultsEvaluation.py C:/Users/matth/OneDrive/Desktop/University/3B/MSCI541-Search-Engines/HW4/hw4-files-2023/results/ C:/Users/matth/OneDrive/Desktop/University/3B/MSCI541-Search-Engines/HW4/hw4-files-2023/qrels/LA-only.trec8-401.450.minus416-423-437-444-447.txt

BM25:

The program accepts three command line arguments: the directory location of your index, the queries file, and the name of a file to store your output

    1. a path to the location of your index, created by IndexEngine

    2. the queries file

    3. the name of a file to store your output

For example, you would run BM25 from the command prompt / terminal / shell as:

    python BM25.py C:/Users/matth/OneDrive/Desktop/University/3B/MSCI541-Search-Engines/HW4/store/term_files queries.txt hw4-bm25-baseline-merxlebe.txt

For this program, the queries file (search topics file) is in the /topics files folder. In the code for the program, the path is hardcoded as:

C:/Users/matth/OneDrive/Desktop/University/3B/MSCI541-Search-Engines/HW4/msci-541-f23-hw4-matterxleben/topics files/" + queries_file

Please change this to your own computers path to the msci-541-f23-hw4-matterxleben/topics files/ folder to obtain the queries (search topics)!

7b)

The results of the baseline BM25 retrieval can be found at hw4-bm25-baseline-merxlebe.txt

```
results > ≡ hw4-bm25-baseline-merxlebe.txt
  1    401 Q0 LA101790-0075 1 6.074748903361086 merxlebeAND
  2    401 Q0 LA021890-0100 2 6.032953328250141 merxlebeAND
  3    401 Q0 LA050690-0109 3 5.6400852726916915 merxlebeAND
  4    401 Q0 LA100890-0131 4 5.630885051378108 merxlebeAND
  5    401 Q0 LA040789-0015 5 5.5405520405981825 merxlebeAND
  6    401 Q0 LA021490-0049 6 5.5367825615584305 merxlebeAND
  7    401 Q0 LA031590-0102 7 5.4670164670752435 merxlebeAND
  8    401 Q0 LA071890-0073 8 5.42381110282183 merxlebeAND
  9    401 Q0 LA040590-0157 9 5.383676256052823 merxlebeAND
 10    401 Q0 LA050790-0042 10 5.379201506182712 merxlebeAND
 11    401 Q0 LA082789-0152 11 5.367573570786622 merxlebeAND
 12    401 Q0 LA050789-0068 12 5.332315133499008 merxlebeAND
 13    401 Q0 LA111089-0188 13 5.32344019428111 merxlebeAND
 14    401 Q0 LA021190-0168 14 5.307385769457557 merxlebeAND
 15    401 Q0 LA060890-0011 15 5.299247923623829 merxlebeAND
 16    401 Q0 LA020789-0133 16 5.229213772444686 merxlebeAND
 17    401 Q0 LA030990-0189 17 5.222206946868615 merxlebeAND
 18    401 Q0 LA050990-0043 18 5.193688047201026 merxlebeAND
 19    401 Q0 LA062290-0172 19 5.183174087943154 merxlebeAND
 20    401 Q0 LA090389-0171 20 5.162111967570969 merxlebeAND
 21    401 Q0 LA050390-0176 21 5.136880574677754 merxlebeAND
 22    401 Q0 LA043090-0043 22 5.1240766775460465 merxlebeAND
 23    401 Q0 LA112189-0066 23 5.12210504543815 merxlebeAND
 24    401 Q0 LA091490-0080 24 5.120933739822092 merxlebeAND
 25    401 Q0 LA061290-0074 25 5.111363185757659 merxlebeAND
 26    401 Q0 LA021590-0209 26 5.110150730361098 merxlebeAND
 27    401 Q0 LA082389-0142 27 5.096443398096712 merxlebeAND
 28    401 Q0 LA010489-0026 28 5.086599738681089 merxlebeAND
 29    401 Q0 LA021590-0208 29 5.070719867194082 merxlebeAND
 30    401 Q0 LA022290-0221 30 5.0381508547565685 merxlebeAND
 31    401 Q0 LA100289-0097 31 5.037974546489284 merxlebeAND
 32    401 Q0 LA041290-0088 32 5.025387408972495 merxlebeAND
 33    401 Q0 LA030990-0070 33 5.021008826321708 merxlebeAND
 34    401 Q0 LA040689-0200 34 5.016233306183249 merxlebeAND
 35    401 Q0 LA030690-0168 35 5.015236790714735 merxlebeAND
 36    401 Q0 LA120489-0110 36 5.011677817493935 merxlebeAND
 37    401 Q0 LA121589-0087 37 5.003802403664816 merxlebeAND
 38    401 Q0 LA100189-0189 38 4.997092130780651 merxlebeAND
 39    401 Q0 LA020690-0078 39 4.996780802615824 merxlebeAND
 40    401 Q0 LA011289-0071 40 4.994797474781549 merxlebeAND
 41    401 Q0 LA031490-0158 41 4.993503766204197 merxlebeAND
 42    401 Q0 LA050889-0081 42 4.9926448470116656 merxlebeAND
 43    401 Q0 LA120689-0034 43 4.988344867167488 merxlebeAND
 44    401 Q0 LA111089-0142 44 4.972261265372243 merxlebeAND
 45    401 Q0 LA100890-0076 45 4.96563088015223 merxlebeAND
 46    401 Q0 LA100190-0033 46 4.959097760451025 merxlebeAND
 47    401 Q0 LA030790-0048 47 4.955181895485336 merxlebeAND
 48    401 Q0 LA061089-0056 48 4.94860218599951 merxlebeAND
 49    401 Q0 LA083090-0247 49 4.9421797218951316 merxlebeAND
 50    401 Q0 LA111289-0073 50 4.935030586177118 merxlebeAND
 51    401 Q0 LA091289-0149 51 4.931766167483419 merxlebeAND
 52    401 Q0 LA101990-0026 52 4.9232339472981 merxlebeAND
 53    401 Q0 LA071790-0033 53 4.912806381325571 merxlebeAND
```

Topic:		AP:		P@10:		NDCG@10:		NDCG@1000:	
Topic:	401	AP:	0.0468	P@10:	0.1	NDCG@10:	0.0948	NDCG@1000:	0.3455
Topic:	402	AP:	0.0607	P@10:	0.2	NDCG@10:	0.2935	NDCG@1000:	0.2668
Topic:	403	AP:	0.5075	P@10:	0.6	NDCG@10:	0.5302	NDCG@1000:	0.7407
Topic:	404	AP:	0.004	P@10:	0	NDCG@10:	0	NDCG@1000:	0.144
Topic:	405	AP:	0.0347	P@10:	0.2	NDCG@10:	0.1518	NDCG@1000:	0.1614
Topic:	406	AP:	0.3941	P@10:	0.2	NDCG@10:	0.3833	NDCG@1000:	0.7362
Topic:	407	AP:	0.2063	P@10:	0.6	NDCG@10:	0.6434	NDCG@1000:	0.5775
Topic:	408	AP:	0.0941	P@10:	0.4	NDCG@10:	0.5389	NDCG@1000:	0.363
Topic:	409	AP:	0.1	P@10:	0.1	NDCG@10:	0.2891	NDCG@1000:	0.2891
Topic:	410	AP:	1	P@10:	0.4	NDCG@10:	1	NDCG@1000:	1
Topic:	411	AP:	0.0983	P@10:	0.2	NDCG@10:	0.3052	NDCG@1000:	0.3759
Topic:	412	AP:	0.3609	P@10:	0.7	NDCG@10:	0.65	NDCG@1000:	0.6379
Topic:	413	AP:	0.0101	P@10:	0	NDCG@10:	0	NDCG@1000:	0.1505
Topic:	414	AP:	0.092	P@10:	0.1	NDCG@10:	0.1672	NDCG@1000:	0.3251
Topic:	415	AP:	0.25	P@10:	0.1	NDCG@10:	0.3904	NDCG@1000:	0.3904
Topic:	417	AP:	0.339	P@10:	0.7	NDCG@10:	0.7885	NDCG@1000:	0.7172
Topic:	418	AP:	0.1394	P@10:	0.6	NDCG@10:	0.7223	NDCG@1000:	0.3433
Topic:	419	AP:	0.5084	P@10:	0.2	NDCG@10:	0.6367	NDCG@1000:	0.7353
Topic:	420	AP:	0.623	P@10:	0.8	NDCG@10:	0.8522	NDCG@1000:	0.8876
Topic:	421	AP:	0.018	P@10:	0	NDCG@10:	0	NDCG@1000:	0.2446
Topic:	422	AP:	0.3624	P@10:	0.4	NDCG@10:	0.3251	NDCG@1000:	0.6452
Topic:	424	AP:	0.0206	P@10:	0.2	NDCG@10:	0.1428	NDCG@1000:	0.1155
Topic:	425	AP:	0.2851	P@10:	0.6	NDCG@10:	0.6316	NDCG@1000:	0.5657
Topic:	426	AP:	0.0274	P@10:	0.3	NDCG@10:	0.2221	NDCG@1000:	0.1698
Topic:	427	AP:	0.0797	P@10:	0.2	NDCG@10:	0.3052	NDCG@1000:	0.3041
Topic:	428	AP:	0.2533	P@10:	0.1	NDCG@10:	0.3904	NDCG@1000:	0.4821
Topic:	429	AP:	0.2827	P@10:	0.1	NDCG@10:	0.3904	NDCG@1000:	0.5748
Topic:	430	AP:	0.4811	P@10:	0.4	NDCG@10:	0.6278	NDCG@1000:	0.6823
Topic:	431	AP:	0.0984	P@10:	0	NDCG@10:	0	NDCG@1000:	0.3367
Topic:	432	AP:	0.0036	P@10:	0	NDCG@10:	0	NDCG@1000:	0.1034
Topic:	433	AP:	0.0052	P@10:	0	NDCG@10:	0	NDCG@1000:	0.1109
Topic:	434	AP:	0.55	P@10:	0.1	NDCG@10:	0.6131	NDCG@1000:	0.7527
Topic:	435	AP:	0.064	P@10:	0.1	NDCG@10:	0.0784	NDCG@1000:	0.3638
Topic:	436	AP:	0.0376	P@10:	0.3	NDCG@10:	0.2528	NDCG@1000:	0.203
Topic:	438	AP:	0.0954	P@10:	0.2	NDCG@10:	0.1357	NDCG@1000:	0.4326
Topic:	439	AP:	0.0038	P@10:	0	NDCG@10:	0	NDCG@1000:	0.1264
Topic:	440	AP:	0.5823	P@10:	0.6	NDCG@10:	0.6937	NDCG@1000:	0.8198
Topic:	441	AP:	0.6079	P@10:	0.6	NDCG@10:	0.7381	NDCG@1000:	0.7381
Topic:	442	AP:	0.0252	P@10:	0.2	NDCG@10:	0.1447	NDCG@1000:	0.2129
Topic:	443	AP:	0.1089	P@10:	0.2	NDCG@10:	0.2173	NDCG@1000:	0.4326
Topic:	445	AP:	0.2444	P@10:	0.2	NDCG@10:	0.4162	NDCG@1000:	0.4162
Topic:	446	AP:	0.0292	P@10:	0	NDCG@10:	0	NDCG@1000:	0.2063
Topic:	448	AP:	0.0175	P@10:	0	NDCG@10:	0	NDCG@1000:	0.2111
Topic:	449	AP:	0.0095	P@10:	0	NDCG@10:	0	NDCG@1000:	0.0969
Topic:	450	AP:	0.2235	P@10:	0.4	NDCG@10:		NDCG@1000:	0.5861

7c)

The results of the baseline BM25 retrieval can be found at hw4-bm25-stem-merxlebe.txt

```
 1   401 Q0 LA030690-0168 1 6.571483823576605 merxlebeAND
 2   401 Q0 LA050690-0109 2 6.460391036324291 merxlebeAND
 3   401 Q0 LA041290-0088 3 6.195006855170041 merxlebeAND
 4   401 Q0 LA040989-0072 4 6.172135707298894 merxlebeAND
 5   401 Q0 LA110690-0075 5 5.705613351577572 merxlebeAND
 6   401 Q0 LA050789-0068 6 5.661913089819003 merxlebeAND
 7   401 Q0 LA100890-0131 7 5.579819203949144 merxlebeAND
 8   401 Q0 LA021490-0049 8 5.571769295008746 merxlebeAND
 9   401 Q0 LA031590-0102 9 5.52464248254968 merxlebeAND
10   401 Q0 LA050790-0042 10 5.51258405708243 merxlebeAND
11   401 Q0 LA071890-0073 11 5.39108117431041 merxlebeAND
12   401 Q0 LA070989-0066 12 5.377856459418849 merxlebeAND
13   401 Q0 LA040590-0157 13 5.333143411638125 merxlebeAND
14   401 Q0 LA111089-0188 14 5.326337522258674 merxlebeAND
15   401 Q0 LA021890-0100 15 5.3156148123297715 merxlebeAND
16   401 Q0 LA021190-0168 16 5.260124850054082 merxlebeAND
17   401 Q0 LA060890-0011 17 5.248969539016584 merxlebeAND
18   401 Q0 LA030990-0189 18 5.228500594961867 merxlebeAND
19   401 Q0 LA112990-0229 19 5.221609137526233 merxlebeAND
20   401 Q0 LA050990-0043 20 5.220355308798698 merxlebeAND
21   401 Q0 LA062789-0019 21 5.218694409415905 merxlebeAND
22   401 Q0 LA061290-0074 22 5.216610031634419 merxlebeAND
23   401 Q0 LA102390-0068 23 5.201782239675228 merxlebeAND
24   401 Q0 LA072590-0049 24 5.201028527677608 merxlebeAND
25   401 Q0 LA091490-0080 25 5.152900208458081 merxlebeAND
26   401 Q0 LA031490-0158 26 5.146041020270798 merxlebeAND
27   401 Q0 LA062290-0172 27 5.1296766207072615 merxlebeAND
28   401 Q0 LA100490-0081 28 5.128459397144992 merxlebeAND
29   401 Q0 LA072790-0172 29 5.1232437985758885 merxlebeAND
30   401 Q0 LA051190-0084 30 5.123202529715591 merxlebeAND
31   401 Q0 LA100289-0097 31 5.0999981800200604 merxlebeAND
32   401 Q0 LA060489-0038 32 5.094759666898801 merxlebeAND
33   401 Q0 LA050390-0176 33 5.091258099252266 merxlebeAND
34   401 Q0 LA043090-0043 34 5.081009980575692 merxlebeAND
35   401 Q0 LA031389-0076 35 5.077286244562838 merxlebeAND
36   401 Q0 LA021590-0209 36 5.068460595146931 merxlebeAND
37   401 Q0 LA082389-0142 37 5.051896332226523 merxlebeAND
38   401 Q0 LA010489-0026 38 5.043525314263908 merxlebeAND
39   401 Q0 LA021490-0144 39 5.043488003913742 merxlebeAND
40   401 Q0 LA021590-0208 40 5.027858461604249 merxlebeAND
41   401 Q0 LA022790-0091 41 5.006450673232987 merxlebeAND
42   401 Q0 LA122190-0057 42 5.002524208025757 merxlebeAND
43   401 Q0 LA121789-0230 43 4.99618587779891 merxlebeAND
44   401 Q0 LA050889-0081 44 4.994612483680944 merxlebeAND
45   401 Q0 LA121589-0087 45 4.9889579658041034 merxlebeAND
46   401 Q0 LA022290-0221 46 4.987029355657999 merxlebeAND
47   401 Q0 LA030990-0070 47 4.979825651025676 merxlebeAND
48   401 Q0 LA011990-0099 48 4.974834551970854 merxlebeAND
49   401 Q0 LA120489-0110 49 4.9682016182297035 merxlebeAND
50   401 Q0 LA110990-0013 50 4.967686878028147 merxlebeAND
51   401 Q0 LA042490-0209 51 4.96231996063267 merxlebeAND
52   401 Q0 LA040689-0200 52 4.958702081628088 merxlebeAND
53   401 Q0 LA100189-0189 53 4.957842202393272 merxlebeAND
54   401 Q0 LA020690-0078 54 4.955621916806248 merxlebeAND
```

Topic	AP:	P@10:	NDCG@10:	NDCG@1000:
401	0.1032	0.3	0.2466	0.45
402	0.2076	0.3	0.3836	0.6017
403	0.5075	0.6	0.5302	0.7407
404	0.0097	0	0	0.1694
405	0.0264	0.1	0.0734	0.1476
406	0.4479	0.3	0.4617	0.77
407	0.1689	0.4	0.5036	0.5224
408	0.136	0.3	0.3571	0.4619
409	0.1	0.1	0.2891	0.2891
410	1	0.4	1	1
411	0.1799	0.3	0.4441	0.4762
412	0.4468	0.8	0.7163	0.7271
413	0.0833	0	0	0.2702
414	0.1055	0.1	0.2021	0.344
415	0.25	0.1	0.3904	0.3904
417	0.3569	0.7	0.7792	0.7333
418	0.2663	0.6	0.7273	0.622
419	0.5833	0.3	0.7542	0.7542
420	0.62	0.8	0.8604	0.8874
421	0.0189	0	0	0.2857
422	0.3756	0.7	0.5559	0.6621
424	0.1544	0.1	0.0636	0.593
425	0.4818	0.7	0.7624	0.8239
426	0.0342	0.1	0.0784	0.1798
427	0.0972	0.2	0.3052	0.3838
428	0.1094	0.1	0.1952	0.3354
429	0.7986	0.4	0.9223	0.9223
430	0.6203	0.4	0.6608	0.7499
431	0.3195	0.6	0.6064	0.6838
432	0.0017	0	0	0.068
433	0.0051	0	0	0.1102
434	0.5417	0.1	0.6131	0.7452
435	0.0411	0	0	0.2813
436	0.0925	0.7	0.7191	0.3182
438	0.1158	0.1	0.0851	0.4579
439	0.0146	0	0	0.1681
440	0.5683	0.5	0.6274	0.8131
441	0.6496	0.6	0.7634	0.7634
442	0.0231	0.2	0.137	0.1976
443	0.1054	0.2	0.1763	0.4203
445	0.2444	0.2	0.4162	0.4162
446	0.0245	0	0	0.2323
448	0.0092	0	0	0.1947
449	0.0075	0	0	0.0913
450	0.2393	0.4	0.4381	0.6497

7d)

After obtaining the BM25 results for the baseline and stem versions, I calculated the four effectiveness measures to evaluate the two results. This can be seen in the following table:

Run	Mean AP	Mean P@10	Mean NDCG@10	Mean NDCG@1000
baseline	0.209	0.253	0.335	0.425
stem	0.251	0.284	0.374	0.487
p-value	0.005167	0.150811	0.100977	0.001687

As we can see, the Porter Stemmer version of the BM25 results has a higher Mean Average Precision, Mean Precision@10, Mean NDCG@10 and Mean NDCG@1000. This shows an overall improvement over the baseline run. I utilized the t-test to compare the statistical significance of the difference between these two results. As we can see, the p-values computed are extremely small for each effectiveness measure. However, if we are to apply the $p < 0.05$ concept to these results, we can see that only Mean Average Precision and Mean NDCG@1000 are statistically significant. Although the other two evaluation metrics are close to 0.05, this test shows that the different ins not statistically significant. This shows that the Porter Stemmer

version is better in overall Mean Average Precision and Mean NDCG@1000. As we can see, these two evaluation metrics have something in common. Both evaluation metrics are evaluating a large sample size (Mean Average Precision is across the entire results set, and NDCG @ 1000 is evaluating at rank 1000), versus the other metrics evaluating a smaller sample size (both at rank 10). This demonstrates that the Porter Stemmer's better results become more statistically significant the bigger sample size is.

To see how on a topic-by-topic basis BM25 is in the baseline and stemmed results, I compared each topic's mean effectiveness measure. I utilized Excel to highlight if the two are equivalent, then the column will equal 0 and highlight yellow. If stemmed results are better than the baseline results, then the column equals 1 and highlights green. If stemmed results are worse than the baseline results, then the column equals -1 and highlights red.

| Baseline | | | | | Stem | | | | | Comparison | | | |
Topic	Mean AP	Mean P@10	Mean NDCG@10	Mean NDCG@1	Topic	Mean AP	Mean P@10	Mean NDCG@10	Mean NDCG@1	Mean AP	Mean P@10	Mean NDCG@10	Mean NDCG@1000
401	0.0468	0.1	0.0948	0.3455	401	0.1032	0.3	0.2466	0.45	1	1	1	1
402	0.0607	0.2	0.2935	0.2668	402	0.2076	0.3	0.3836	0.6017	1	1	1	1
403	0.5075	0.6	0.5302	0.7407	403	0.5075	0.6	0.5302	0.7407	0	0	0	0
404	0.004	0	0	0.144	404	0.0097	0	0	0.1694	1	0	0	1
405	0.0347	0.2	0.1518	0.1614	405	0.0264	0.1	0.0734	0.1476	-1	-1	-1	-1
406	0.3941	0.2	0.3833	0.7362	406	0.4479	0.3	0.4617	0.77	1	1	1	1
407	0.2063	0.6	0.6434	0.5775	407	0.1689	0.4	0.5036	0.5224	-1	-1	-1	-1
408	0.0941	0.4	0.5389	0.363	408	0.136	0.3	0.3571	0.4619	1	-1	-1	1
409	0.1	0.1	0.2891	0.2891	409	0.1	0.1	0.2891	0.2891	0	0	0	0
410	1	0.4	1	1	410	1	0.4	1	1	0	0	0	0
411	0.0983	0.2	0.3052	0.3759	411	0.1799	0.3	0.4441	0.4762	1	1	1	1
412	0.3609	0.7	0.65	0.6379	412	0.4468	0.8	0.7163	0.7271	1	1	1	1
413	0.0101	0	0	0.1505	413	0.0833	0	0	0.2702	1	0	0	1
414	0.092	0.1	0.1672	0.3251	414	0.1055	0.1	0.2021	0.344	1	0	1	1
415	0.25	0.1	0.3904	0.3904	415	0.25	0.1	0.3904	0.3904	0	0	0	0
417	0.339	0.7	0.7885	0.7172	417	0.3569	0.7	0.7792	0.7333	1	0	-1	1
418	0.1394	0.6	0.7223	0.3433	418	0.2663	0.6	0.7273	0.622	1	0	1	1
419	0.5084	0.2	0.6367	0.7353	419	0.5833	0.3	0.7542	0.7542	1	1	1	1
420	0.623	0.8	0.8522	0.8876	420	0.62	0.8	0.8604	0.8874	-1	0	1	-1
421	0.018	0	0	0.2446	421	0.0189	0	0	0.2857	1	0	0	1
422	0.3624	0.4	0.3251	0.6452	422	0.3756	0.7	0.5559	0.6621	1	1	1	1
424	0.0206	0.2	0.1428	0.1155	424	0.1544	0.1	0.0636	0.593	1	-1	-1	1
425	0.2851	0.6	0.6316	0.5657	425	0.4818	0.7	0.7624	0.8239	1	1	1	1
426	0.0274	0.3	0.2221	0.1698	426	0.0342	0.1	0.0784	0.1798	1	-1	-1	1
427	0.0797	0.2	0.3052	0.3041	427	0.0972	0.2	0.3052	0.3838	1	0	0	1
428	0.2533	0.1	0.3904	0.4821	428	0.1094	0.1	0.1952	0.3354	-1	0	-1	-1
429	0.2827	0.1	0.3904	0.5748	429	0.7986	0.4	0.9223	0.9223	1	1	1	1
430	0.4811	0.4	0.6278	0.6823	430	0.6203	0.4	0.6608	0.7499	1	0	1	1
431	0.0984	0	0	0.3367	431	0.3195	0.6	0.6064	0.6838	1	1	1	1
432	0.0036	0	0	0.1034	432	0.0017	0	0	0.068	-1	0	0	-1
433	0.0052	0	0	0.1109	433	0.0051	0	0	0.1102	-1	0	0	-1
434	0.55	0.1	0.6131	0.7527	434	0.5417	0.1	0.6131	0.7452	-1	0	0	-1
435	0.064	0.1	0.0784	0.3638	435	0.0411	0	0	0.2813	-1	-1	-1	-1
436	0.0376	0.3	0.2528	0.203	436	0.0925	0.7	0.7191	0.3182	1	1	1	1
438	0.0954	0.2	0.1357	0.4326	438	0.1158	0.1	0.0851	0.4579	1	-1	-1	1
439	0.0038	0	0	0.1264	439	0.0146	0	0	0.1681	1	0	0	1
440	0.5823	0.6	0.6937	0.8198	440	0.5683	0.5	0.6274	0.8131	-1	-1	-1	-1
441	0.6079	0.6	0.7381	0.7381	441	0.6496	0.6	0.7634	0.7634	1	0	1	1
442	0.0252	0.2	0.1447	0.2129	442	0.0231	0.2	0.137	0.1976	-1	0	-1	-1
443	0.1089	0.2	0.2173	0.4326	443	0.1054	0.2	0.1763	0.4203	-1	0	-1	-1
445	0.2444	0.2	0.4162	0.4162	445	0.2444	0.2	0.4162	0.4162	0	0	0	0
446	0.0292	0	0	0.2063	446	0.0245	0	0	0.2323	-1	0	0	1
448	0.0175	0	0	0.2111	448	0.0092	0	0	0.1947	-1	0	0	-1
449	0.0095	0	0	0.0969	449	0.0075	0	0	0.0913	-1	0	0	-1
450	0.2235	0.4	0.3031	0.5861	450	0.2393	0.4	0.4381	0.6497	1	0	1	1

As we can see, topics: 405, 407, 420, 428, 432, 433, 434, 435, 440, 442, 443, 446, 448, 449 are overall equal or better in the baseline than the stemmed results. When looking at these queries, it seems like queries using plurals and -ing endings are more common amongst these results.

This shows that when the user makes a search with these additions to their words, it may be more likely that the results lead to non-relevant documents being returned.