MSCI 541 HW4 Report

Ramandeep Farmaha 20516974

Problem 1

Two documents can have the same retrieval score and be relevant and non-relevant respectively if one of the documents contains a higher count of a more important term in the query than the other document. For example, for the query "President Lincoln", the term "President" appears in 250,000 documents, while "Lincoln" appears in 500. When a user submits the query, "Lincoln" will have a much higher idf weight, since it occurs far fewer than "President". As a result, if two documents a returned, one document that is about the Lincoln Town Car might repeat the word "Lincoln" much more frequently than another document that is an article about President Lincoln. As a result, the first article, which is completely irrelevant to the query, may achieve the same BM25 retrieval score as the relevant article about President Lincoln.

Problem 2

Part A

Long documents can have words that occur once and words (such as "the") appear hundreds of times. Although normalization may stymie the effects of the more frequent terms, it is better to use the logarithm of the number of term occurrences.

Part B

The term frequency of document is represented in BM-25 via the component: `[(k_1 + 1)f_i]/(K + f_i)`. The `f_i` term is the frequency of term i in the document, while `k_1` is a constant. `K = k_1[(1 - b) +b * dl/avdl]`, where `b` is another constant, `dl` and `avdl` are the document length and average document length of all documents respectively. Thus, the term frequency, represented in the numerator, is normalized by the length of the document (the `K` variable) in the denominator. This achieves a similar effect to computing the logarithm of the term frequency in the document, which is used in the tf-idf calculation.

Problem 3

Stemming would speed up the retrieval speed, as there would be fewer terms indexed, producing a much thinner vocabulary. For example, the words: "fish", "fishes", "fishing" would all be stemmed to "fish", thus reducing the number of words that involve fish from 3 to 1 in the vocabulary. When a query for "fishes" appears, its stem (i.e. "fish") would be retrieved from the vocabulary.

Problem 4

Words in queries that are not contained in the document collection are ignored by the retrieval model. In practice, this means that the model assigns a score of 0 for the query term that doesn't exist in the document collection.

Problem 5

```
Total document length = 131,896
Total vocabulary size = 247,031

Matrix = 131896 * 247031 = 32,582,400,776 cells
```

If each cell is 4 bytes, this would take up a total memory space of:

```
    memory_size = 130,329,603,104 bytes = 121.38 GB
```

There are a total of `31,916,824` doc_ids for postings in the postings list. Thus,

```
Empty cells = 32,582,400,776 - 31,916,824 = 32,550,483,952
memory_savings = 32,550,483,952*4 = 130,201,935,808 bytes = 121.26 GB
```

Problem 6

The largest source of error in Gary's design is the fact that he chose to remove stopwords and stem tokens using the Porter stemmer for queries, but not for documents in the collection. The stemmed and stopword-removed tokens would then be evaluated against the document collection. For example, if the query is "Germany boats", after stemming, it becomes "Germani, boat", however the document collection may not contain the word "Germani" because it hasn't been stemmed. This would greatly reduce the performance of the retrieval system.

Problem 7

Part D

A: Performance of Baseline vs Porter Stemmer

Run Name	Mean Average Precision	Mean P@10	Mean NDCG@10	Mean NDCG@1000	Mean TBG
rfarmaha-hw4-bm25-stem	0.251	0.284	0.374	0.486	2.047
rfarmaha-hw4-bm25-baseline	0.208	0.251	0.333	0.425	1.766

The results above indicate that all evaluation measures experienced significant improvement from the baseline after performing stemming.

B: Statistical Significance Tests

Effectiveness Measure	Best Run Score	Second Best Run Score	Relative Percent Improvement	Two-sided Paired t-Test p-value
Mean Average Precision	0.251	0.208	20.673%	0.005181800894604008

Effectiveness Measure	Best Run Score	Second Best Run Score	Relative Percent Improvement	Two-sided Paired t-Test p-value
Mean P@10	0.284	0.251	13.147%	0.1251062406735981
Mean NDCG@10	0.374	0.333	12.312%	0.09021119902830947
Mean NDCG@1000	0.486	0.425	14.353%	0.001618024355585357
Mean TBG	2.047	1.766	15.912%	0.0005389512456214771

NOTE: The stemmed retrieval system was the best run score for all metrics in the table above.

From the p-values above, we can determine that the improvements to Mean Average Precision, Mean NDCG@1000, and Mean TBG were statistically significant, since their p-values scored below 0.05.

C: Per-topic Comparison of Performance

To perform a per-topic comparison of performance between baseline and stemming, we can analyze the Average Precision scores for each topic:

topic_id	baseline	porter_stemmer	difference
401	0.04520136245167289	0.10411055698541759	0.0589091945337447
402	0.060583178844830304	0.20679423147331863	0.14621105262848832
403	0.5075373854785619	0.5075373854785619	0.0
404	0.004080854309687262	0.009737076648841355	0.005656222339154093
405	0.031189797856464523	0.02639092389092389	-0.004798873965540632
406	0.3937300990779514	0.43962551609610434	0.04589541701815292
407	0.2049215955326435	0.16781521664312588	-0.03710637888951762
408	0.09428896657465555	0.13602167770378476	0.04173271112912921
409	0.1	0.1	0.0
410	1.0	1.0	0.0
411	0.09232970514432488	0.1775343846381666	0.08520467949384171
412	0.36413385360822703	0.45429034350192804	0.09015648989370101
413	0.0101010101010102	0.08333333333333333	0.07323232323232323
414	0.09199280251911829	0.10541979949874687	0.013426996979628583
415	0.25	0.25	0.0
417	0.3372253000175782	0.35537142472554184	0.018146124707963618
418	0.13859837561741611	0.263782026110302	0.12518365049288588
419	0.5084805739945334	0.5833333333333334	0.07485275933879998
420	0.6216943897542843	0.617050353172872	-0.004644036581412214
421	0.017413998299527823	0.018931557573567133	0.0015175592740393103
422	0.36528367227152864	0.37623030016503206	0.010946627893503424
424	0.020552056737392257	0.15405647887569035	0.1335044221382981
425	0.2847815216801742	0.48215491228549373	0.19737339060531955
426	0.02755211103337166	0.03446992861728693	0.006917817583915265
427	0.07982614016051345	0.09740455460942925	0.017578414448915794
428	0.25336107074913494	0.10812841530054644	-0.1452326554485885

topic_id	baseline	porter_stemmer	difference
429	0.28126786093374917	0.7986111111111112	0.517343250177362
430	0.4811325611325611	0.6202564102564103	0.13912384912384923
431	0.0947311875159835	0.31724427564866625	0.22251308813268275
432	0.003531948268021692	0.0016822788852686572	-0.001849669382753035
433	0.005198412698412699	0.005032206119162641	-0.00016620657925005798
434	0.55	0.5434782608695652	-0.006521739130434856
435	0.06252079842333978	0.03947754839963145	-0.023043250023708335
436	0.03744550774375304	0.0896204315407874	0.05217492379703436
438	0.09454928239552642	0.11505232965550223	0.020503047259975815
439	0.0037277440919837924	0.014590316027476154	0.010862571935492362
440	0.5822984052355868	0.5682617814773546	-0.014036623758232225
441	0.6079365079365079	0.6496031746031746	0.04166666666666674
442	0.02486840999715128	0.022895624473549328	-0.001972785523601951
443	0.1088188445075541	0.10541211947233003	-0.003406725035224059
445	0.24444444444444446	0.24444444444444446	0.0
446	0.029155490169778902	0.024978826146836566	-0.004176664022942336
448	0.01759230594354984	0.009247231086586797	-0.008345074856963045
449	0.00950168918918919	0.007454867827208253	-0.002046821361980936
450	0.22565571592103092	0.24360351977412997	0.017947803853099048

For the majority of the topics, the Porter Stemmer improved the average precision of the retrieval system. The topics below are those that achieved the same average precision rates before and after stemming (i.e. they had a difference = 0) along with a brief explanation:

topic_id	explanation
403	Osteoperosis is a unique word that probably doesn't have a stemmed version
409	"legal", "Pan", "Am" "103" all seem like terms that cannot be further stemmed
410	Schengen seems like a name, thus requiring an exact match
415	Common words "golden" and "triangle" together represent a name semantically ("Golden Triangle")
445	"Women" and "clergy" may have no further stemmed forms

There are also several topics that had lower average precision scores than the baseline. This may be due to loss of contextual information after stemming was done to the tokens in the query and documents.