CS 510 Experimental Lab

Daniel Campos

April 14th,2021

1 General Info

Our team is blender, our team is: Daniel Campos and our task is the 2021 CLEF eHealth IR Task on consumer health search: Subtask 1

2 Goal of Participation

The goal of participation in this task is to build real world experience on how to create a high quality search stack on a novel corpus. Using the health search corpus I explored the effect that variations in ranking methods can have on model performance. I seek to understand how variations in BM25 and RM3 hyperparameters affect relevance. By understanding how variations in BM25 hyperparameters effect relevance, authority, and readability.

3 Activities

Our activities broadly can be categorized into three areas: index generation, search system creation, and model experimentation. For index generation we build scripts which take the CLEF corpus and turn it into a format accessible to the PySerini search library. Using this processed corpus we then generate the index which we use for our experiments. Index creation generally was quite slow but in experimentation's on using more threads we were able to have large speedups. Once we had an index we went ahead and created our search systems. Leveraging the PySerini library, we created search methodologies which would allow various forms of experimentation around sparse search hyper parameters and on combinations with dense search methods. The drawbacks in our experiments area mostly have to do with corpus size. Since the corpus is enormous (about 500 gigabytes) any failure is cascading and means discovery of minor issues (naming format) require starting from scratch and take days to run. We wish we could say that it only took us one attempt but in reality we had many mistakes which caused us to repeat slow processes like indexing and document processing. Once we finalized creating search systems we began to produce runs varying all parameters. All of our code can be found on our GitHub repository [1] and our index is available in blob storage [2]

[1] https://github.com/spacemanidol/CS510IR/tree/main/Competition
[2] https://spacemanidol.blob.core.windows.net/blob/clef2018index.tar.gz

4 Findings

The CLEF eHealth ranking task is centered around a document corpus leveraging CLUEWeb which features 1903 domains and 5.5 million URLS. While our experiments were focused on the 2021 queries, their evaluation is not yet available so we base our findings on the 2018 task. The 2018 task, like the 2021 task features 50 queries issued on the Health on the Net search service. The rankings returned are evaluated in TREC style and there are relevance files optimizing to the document retrievals relevance, authority/trust, and readability. While theoretically evaluation can happen across all TREC metrics there is a focus on NDCG@10, BPref, and RBP. For our experiments we focus only on BM25. We perform 3 sets of experiments in varying the BM25 ranker, varying the hyperparameter for K, varying the hyperparameter for B, and varying the importance of the original query when using query rewriting system RM3. We set the BM25 hyperparameters to standard values of 0.75 for B, 1.2 for K and 1.0 for query term importance.
To explore the effects of varying B we fix all hyperparameters on a simple BM25 ranker and gradually vary values of B by 0.1 [0.0, 1.0]. As shown in 1 we see that when focused on relevance, a higher value of B is better but as it passes a threshold (somewhere around 0.9) performance decreases. This is slightly higher than the expected and recommended value of 0.75 but we attribute this to the medical nature of these queries. Higher b values driving improved performance generally holds for readability and trust but these other metrics do not have much more condensed variation. We do not see large changes in readability scores despite variation in values which leads us to believe that B does not interact with terms that can represent readability.
To explore the effects of varying the value for K we once again fix other values and gradually increase the value of K by 0.1 for [1.0,2]. As shown in 2 we see relevance is relatively uncharged despite variations in K. This is surprising because K values can usually be optimized for the task yet in our results we see no such effect.
To explore the effect of varying the importance of the original query term we fix BM25 hyperparamets and gradually vary values for initial query term importance by 0.1 [0.0, 1.0]. For query rewriting we explore the simple/standard RM3 engine with a fixed term expansion of 10 terms and 10 document expansions. Given the domain we are not surprised to find that are shown in 3. Medical queries arguably have a much lower fault tolerance for dropped query terms and as a result query variation causes substantial drops. At a high level, we observe that the the average NDCG is quite low despite there being a huge variation in NDCG per query. We attribute this to the difficulty in health search as some are general keyword like: do allergies cause migraines while others use complex and expert language like: antiandrogen therapy for prostate cancer. We believe that given BM25s low performance across the board a second stage ranking engine would be beneficial as we see the recall at 100 scores in the mid 50s across the board. Our experiments point us to the hypothesis that a BM25 ranker with a B of , a K of , and a original query importance of are optimal. We believes this is the case because of what we discussed for query rewriting (sensitivity of medical queries to term dropping) and because how well these rules of thumbs for search are.

5 Error Analysis

To explore some specific examples of failure in query rewriting we looked at individual queries. We find that the query "health benefits of spiraling" vary from 0.0552 to 0.1663 depending on original query importance (0.1, 1). Looking at the retrieved documents we can understand that any

B value	relevance	readability	trust
0.0	0.0583	0.0425	0.0486
0.1	0.0583	0.0425	0.0486
0.2	0.0662	0.0439	0.0516
0.3	0.0691	0.0445	0.0524
0.4	0.0719	0.0458	0.0540
0.5	0.0736	0.0463	0.0543
0.6	0.0739	0.0465	0.0536
0.7	0.0750	0.0471	0.0535
0.8	0.0782	0.0480	0.0530
0.9	0.0777	0.0473	0.0521
1.0	0.0626	0.0368	0.0403

Table 1: Results for variation in BM25 B value

v	relevance	readability	trust
1.0	0.0053	0.0049	0.0053
1.1	0.0720	0.0465	0.0539
1.2	0.0719	0.0465	0.0538
1.3	0.0719	0.0461	0.0538
1.4	0.0719	0.0461	0.0540
1.5	0.0721	0.0463	0.0541
1.6	0.0720	0.0463	0.0541
1.7	0.0722	0.0464	0.0543
1.8	0.0723	0.0464	0.0542
1.9	0.0727	0.0466	0.0542
2.0	0.0621	0.0466	0.0542

Table 2:

Original Query Importance	relevance	readability	trust
0.0	0.0487	0.0312	0.0366
0.1	0.0537	0.0337	0.0398
0.2	0.0560	0.0357	0.0415
0.3	0.0585	0.0374	0.0430
0.4	0.0597	0.0378	0.0432
0.5	0.0599	0.0391	0.0443
0.6	0.0612	0.0391	0.0451
0.7	0.0626	0.0399	0.0461
0.8	0.0639	0.0406	0.0466
0.9	0.0680	0.0426	0.0494
1.0	0.0756	0.0475	0.0531

Table 3: Results for variation in BM25 scores with variations in in initial query importance

rewriting or term modification can greatly change the intent which makes it an unideal candidate for rewriting. While simple rm3 does not show itself to be robust with medical queries we believe methods that can leverage semantic similarity like word2vec would likely provide a better query rewriting method because they maintain relative query information.

6 Summary

In this experimental lab what we have learned how a performant search engine can be setup on a novel search task. As shown in our results, we found that there are large variations in relevance based on minor changes in ranking methods, query rewriting as implemented with RM3 is not robust for medical queries, and the general guidelines for k and b values are effective and relevant across tasks.

7 Task division

Since Daniel is the only member of this group he has done everything.