Assignment 4

Algorithm of Solution

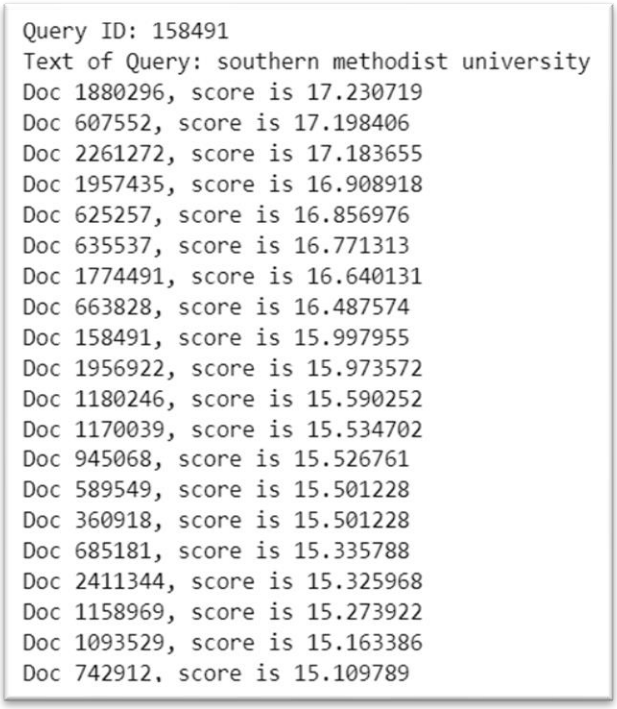
To begin, we established the Elastic Search configuration. Next, we proceeded to read the "*documents*" from the wikiIR collection and indexed them using the pre-configured Elastic Search setup. Subsequently, we read a "*test/queries*" document and generated a list of the top 20 BM25 scores for each of the 100 documents in the testing document. Finally, we evaluated the results by comparing the generated document with the "*testing/qrels*" document.

We utilized pre-existing documents that had been scored with the BM25 algorithm and applied the *SentenceTransformer* model ("*msmarco-distilbert-cos-v5*") to them. This model uses an encoder to convert sentences into embeddings, which can be used to calculate similarity scores between them. We then used the *utils.cos sim* function to compute cosine similarity scores for each query-document pair in the set, and re-evaluated the scores based on the new similarity values obtained from the SentenceTransformer model.

Top 20 documents scores

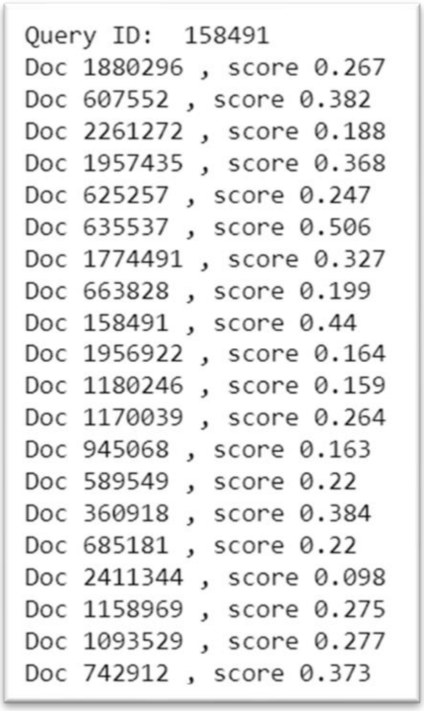
For proof of concept here will be shown the top 20 results for the first query (158491).

BM25 Scoring Document



```
Query ID: 158491
Text of Query: southern methodist university
Doc 1880296, score is 17.230719
Doc 607552, score is 17.198406
Doc 2261272, score is 17.183655
Doc 1957435, score is 16.908918
Doc 625257, score is 16.856976
Doc 635537, score is 16.771313
Doc 1774491, score is 16.640131
Doc 663828, score is 16.487574
Doc 158491, score is 15.997955
Doc 1956922, score is 15.973572
Doc 1180246, score is 15.590252
Doc 1170039, score is 15.534702
Doc 945068, score is 15.526761
Doc 589549, score is 15.501228
Doc 360918, score is 15.501228
Doc 685181, score is 15.335788
Doc 2411344, score is 15.325968
Doc 1158969, score is 15.273922
Doc 1093529, score is 15.163386
Doc 742912. score is 15.109789
```

Cosine Similarity Scoring Document



```
Query ID:  158491
Doc 1880296 , score 0.267
Doc 607552 , score 0.382
Doc 2261272 , score 0.188
Doc 1957435 , score 0.368
Doc 625257 , score 0.247
Doc 635537 , score 0.506
Doc 1774491 , score 0.327
Doc 663828 , score 0.199
Doc 158491 , score 0.44
Doc 1956922 , score 0.164
Doc 1180246 , score 0.159
Doc 1170039 , score 0.264
Doc 945068 , score 0.163
Doc 589549 , score 0.22
Doc 360918 , score 0.384
Doc 685181 , score 0.22
Doc 2411344 , score 0.098
Doc 1158969 , score 0.275
Doc 1093529 , score 0.277
Doc 742912 , score 0.373
```

Total documents number:10000

Sentence transformer model: 'msmarco-distilbert-cos-v5'

Evaluation of BM25 and Cosine Similarity

	p@10	**MAP**	**p@20**
BM25(Top 20)	0.206	0.146	0.148
Cosine Similarity (Top20)	0.234	0.170	0.148

Explanation of scores: By utilizing the cosine similarity function, the MAP and p@10 scores were enhanced, while the p@20 score remained the same due to the evaluation of the identical top 20 documents.

Bonus Question

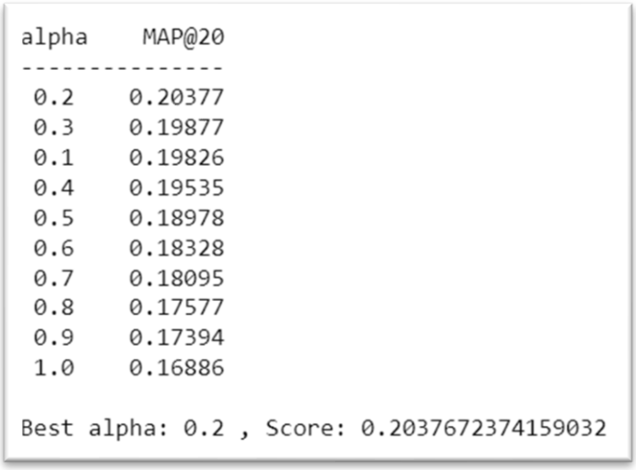
To answer this question, we began by selecting a set of initial "training/queries" queries to create a document of 100 results, with 50 top doc scores from each query. We then computed BM25 scores for these 50 top documents and normalized them to a range of (0,1). Using these normalized BM25 scores, we computed the cosine similarity scores, and finally, we combined the BM25 and cosine similarity results using the formula:

*alpha*BM25 + (1-alpha) * q_d_cosine_similarity*

Through this formula we determined the optimal value for alpha, which maximized the @MAP20 metric, and applied this value to the test query dataset. Overall, this methodology allowed us to effectively combine BM25 and cosine similarity scores to improve retrieval performance on both the training and test datasets.

10 values of alpha that were taken for the training queries were.

[0.1, 0.2, 0,3, 0,4, 0,5, 0.6, 0,7, 0,8, 0.9, 1.0]



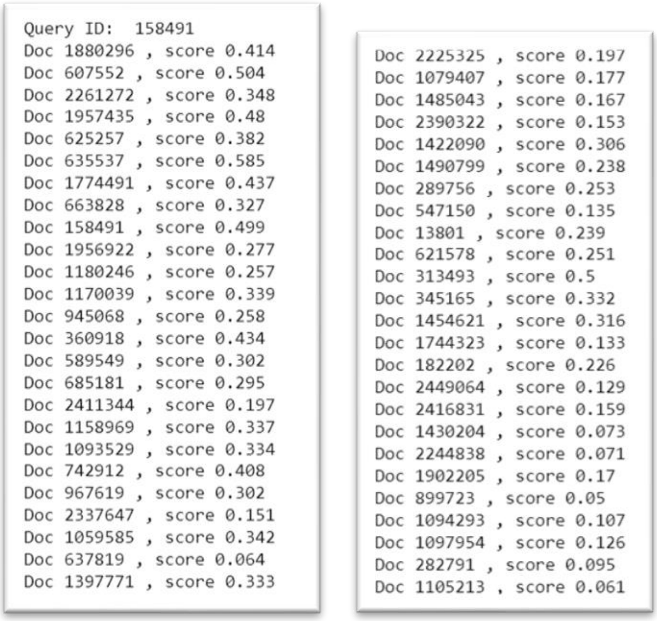
```
alpha    MAP@20
---------------
 0.2    0.20377
 0.3    0.19877
 0.1    0.19826
 0.4    0.19535
 0.5    0.18978
 0.6    0.18328
 0.7    0.18095
 0.8    0.17577
 0.9    0.17394
 1.0    0.16886

Best alpha: 0.2 , Score: 0.2037672374159032
```

The Test Query Evaluation

For the test query, the alpha value that was taken was 0.2, because it maximizes the value of MAP20 for the training query set.

For proof of concept here will be shown the top 50 results for the first query (158491).

BM25+CosineSimilarity(*alpha*=0.2)



```
Query ID:  158491
Doc 1880296 , score 0.414
Doc 607552 , score 0.504
Doc 2261272 , score 0.348
Doc 1957435 , score 0.48
Doc 625257 , score 0.382
Doc 635537 , score 0.585
Doc 1774491 , score 0.437
Doc 663828 , score 0.327
Doc 158491 , score 0.499
Doc 1956922 , score 0.277
Doc 1180246 , score 0.257
Doc 1170039 , score 0.339
Doc 945068 , score 0.258
Doc 360918 , score 0.434
Doc 589549 , score 0.302
Doc 685181 , score 0.295
Doc 2411344 , score 0.197
Doc 1158969 , score 0.337
Doc 1093529 , score 0.334
Doc 742912 , score 0.408
Doc 967619 , score 0.302
Doc 2337647 , score 0.151
Doc 1059585 , score 0.342
Doc 637819 , score 0.064
Doc 1397771 , score 0.333
```

```
Doc 2225325 , score 0.197
Doc 1079407 , score 0.177
Doc 1485043 , score 0.167
Doc 2390322 , score 0.153
Doc 1422090 , score 0.306
Doc 1490799 , score 0.238
Doc 289756 , score 0.253
Doc 547150 , score 0.135
Doc 13801 , score 0.239
Doc 621578 , score 0.251
Doc 313493 , score 0.5
Doc 345165 , score 0.332
Doc 1454621 , score 0.316
Doc 1744323 , score 0.133
Doc 182202 , score 0.226
Doc 2449064 , score 0.129
Doc 2416831 , score 0.159
Doc 1430204 , score 0.073
Doc 2244838 , score 0.071
Doc 1902205 , score 0.17
Doc 899723 , score 0.05
Doc 1094293 , score 0.107
Doc 1097954 , score 0.126
Doc 282791 , score 0.095
Doc 1105213 , score 0.061
```

Testing	**p@10**	**MAP**	**p@20**
BM25	0.206	0.145	0.147
CosineSimilarity	0.233	0.176	0.158
BM25	0.242	0.181	0.164

Conclusion: our experiments have shown that giving more weight to the cosine similarity scores leads to better performance compared to solely relying on BM25 scores, with a split of approximately 80% to 20%, respectively. Additionally, we found that using only BM25 scores with a weight of α = 1 yielded the worst performance out of all values in the [0, 1] range.

Code

GitHub link - https://github.com/ZeroNegativity/ElasticSearch_with_Cosine-Similarity-scoring.git