MSCI 541: HW 3

Jonathan Chen
20722167
November 9th, 2021

General

All programs were created using Java (version 17, 2021-09-14 LTS, Java(TM) SE Runtime Environment (build 17+35-LTS-2724), using the IntelliJ IDE.

To run the programs, the pre-built jar files can be used.

- To run the IndexEngine, go to the *out/artifacts/IndexEngine_jar/* directory in the command line and then run *java -jar IndexEngine.jar [path_to_latimes.gz] [output_path]*.
- To run the GetDoc, go to the *out/artifacts/GetDoc_jar/* directory in the command line and then run *java -jar GetDoc.jar [path to indexed data] ['id' or 'docno'] [id or docno value]*.
- To run the BooleanAND program, go to the *out/artifacts/BooleanAND_jar/* directory in the command line and then run *java -jar BooleanAND.jar [path to indexed data] [path to queries file] [output filename]*.
- To run the Evaluate program, go to the *out/artifacts/Evaluate_jar/* directory in the command line and then run *java -jar Evaluate.jar [path to indexed data] [path to qrels file] [path to results file] [output directory]*

Question 1

A scenario when you'd want to use precision at rank 10 instead of average precision is when the total number of relevant documents is unknown. Oftentimes, it is difficult to quantify the total number of relevant documents, so using precision at rank 10 is desirable because it standardizes the comparison that can be made across different algorithms. As a result, the precision at rank 10 is applicable for any system where you want the most relevant results in the top 10, as opposed to later on. For example, a web search engine (Google) would care about the precision at rank 10 more because the majority of users do not bother to click on the second page of Google search results.

Question 2

The first advantage that nDCG@10 has over precision at rank 10 for the evaluation of a web search engine is that nDCG has a graded relevance. nDCG is commonly graded on a scale of 0 to 5 (from non-relevant to perfect) to determine the relevancy of a document, while precision at rank 10 is a binary relevance of whether the document is relevant or not-relevant. This provides nDCG with a greater level of granularity and is advantageous for web search engines because

documents that are only somewhat related can still be returned by the search engine, as opposed to not returned at all.

The second advantage is that nDCG returns a greater score for the order at which the more relevant documents are returned. In nDCG, search results that return the documents with the greatest gain first are scored higher than those that return results with a lower gain first. On the other hand, precision at rank 10 only considers if a relevant document is returned within the top 10, irrespective of the order at which it was returned. As a result, nDCG is advantageous because it promotes systems that return more relevant documents at the beginning of the search results.

Question 3

a) The randomization and bootstrap methods are appropriate for determining the statistical significance of the difference in the medians.

b) We should use a non-paired test of statistical significance because two different sets of participants were used to study the two systems.

c) To conduct the experiment to allow you to use a paired test, the experiment would need to be designed such that the variability would be reduced between the systems. As a result, one method would be to use the same 50 human participants to test the two systems because this would reduce the variability between the systems that occurs from the unique thoughts and beliefs of each test subject.

d) If we obtain a large p-value of 0.8, we say that we failed to reject the null hypothesis. The idea of the null hypothesis is to find the probability that the sample result would occur, with the end goal of proving the alternative hypothesis. Thus, a high p-value indicates that there is a lack of evidence to suggest that we should accept the alternative hypothesis (or reject the null hypothesis). It does not, however, prove that the null hypothesis should be accepted because the p-value does not provide any evidence of the veracity of the null hypothesis. It is still possible that the null hypothesis should be rejected even though the sample returned a p-value of 0.8 because the sample that was conducted might have simply missed this evidence.

Question 4

a) The p-value of 0.06 means that there is a 6% chance that the observed difference in means would occur if algorithms A and B are treated as identical. In terms of the 1000 samples, 60 of them would contain an absolute difference in means greater than or equal to the difference of 0.18 that was observed in this experiment.

b) My recommended course of action would be to compare algorithms B and C. If the p-value of the mean difference between B and C is less than or equal to 0.05, then I would recommend

algorithm B. Although there isn't a statistically significant improvement of B compared to A at a 95% confidence level, there is still an overall improvement and a statistically significant improvement at a lower confidence level. Moreover, the statistically significant improvement of B compared to C, coupled with the large improvement in the mean nDCG (from 0.01 between A and C to 0.18 between A and B), make up for the slight decrease in the probability that B is statistically significantly better than A. On the other hand, if the p-value is greater than 0.05, then I would recommend algorithm C because algorithm B was not a statistically significant improvement over either of the other algorithms.

Question 5

a) and b)

Run Name	Mean Average Precision	Mean P@10	Mean NDCG@10	Mean NDCG@1000	Mean TBG
student1	**0.250**	**0.282**	**0.371**	**0.485**	**2.037**
student2	0.141	0.193	0.251	0.344	1.251
student3	0.099	0.158	0.181	0.312	1.265
student4	0.202	0.242	0.327	0.427	1.759
student5	*0.224*	0.256	0.320	*0.464*	*1.978*
student6	bad format	bad format	bad format	bad format	bad format
student7	bad format	bad format	bad format	bad format	bad format
student8	0.213	*0.260*	*0.346*	0.438	1.868
student9	0.139	0.204	0.241	0.327	1.590
student10	bad format	bad format	bad format	bad format	bad format
student11	0.137	0.167	0.210	0.299	1.131
student12	bad format	bad format	bad format	bad format	bad format
student13	0.073	0.093	0.115	0.201	0.770
student14	0.200	0.251	0.323	0.415	1.760
msmuckerAND	0.090	0.124	0.211	0.272	0.758

c)
The best run was Student 1's scores. The second best run was a mixture of Student 5 and Student 8's scores. The results of a student's t-test are available in the t-test.csv file.

Effectiveness Measure	Best Run Score	Second Best Run Score	Relative Percent Improvement	Student's t-test, two-side, paired, p-value
Mean AP	0.25	0.224	11.607%	0.171
Mean P@10	0.282	0.260	8.462%	0.243
Mean NDCG@10	0.371	0.346	7.225%	0.248
Mean NDCG@1000	0.485	0.464	4.526%	0.194
Mean TBG	2.037	1.978	2.983%	0.528

d) No best run of a measure was found to be a statistically significant improvement over the second best run because the p value was not less than 0.05 for any of the measures.

e)
For Student 2:
Console output:

```
Successfully calculated the effectiveness measures for student2
Mean Average Precision: 0.14083819107051973
Mean Precision @10: 0.19333333333333325
Mean NDCG @10: 0.25107241022631555
Mean NDCG @1000: 0.3435202024132736
Mean TBG: 1.2513266214830443
```

Files updated/created:
Updated: hw3-5a-jhhchen.csv

```
student2,0.141,0.193,0.251,0.344,1.251,
```

Created: student2.csv

Average Precision,401,0.0403377583185201
Average Precision,402,0.15554315743067199
Average Precision,403,0.5181658314928408

Average Precision,404,0.026792114695340503
Average Precision,405,0.023218294051627383
Average Precision,406,0.5396358524344804
Average Precision,407,0.12691027321387646
Average Precision,408,0.1761361146666389
Average Precision,409,0.07142857142857142
Average Precision,410,0.7028846153846153
Average Precision,411,0.2835203570626717
Average Precision,412,0.0969455240957383
Average Precision,413,0.005405405405405406
Average Precision,414,0.10833333333333334
Average Precision,415,0.125
Average Precision,417,0.05884848769805482
Average Precision,418,0.07067448933608388
Average Precision,419,0.28407014979905004
Average Precision,420,0.48257872961484244
Average Precision,421,0.005804936852296678
Average Precision,422,0.03867659574599077
Average Precision,424,0.05506923888302862
Average Precision,425,0.2720934005508434
Average Precision,426,0.018594560268947468
Average Precision,427,0.05366500273711303
Average Precision,428,0.0111111111111111
Average Precision,429,0.25
Average Precision,430,0.3990972950304047
Average Precision,431,0.1421563816876285
Average Precision,432,0.0026239052263011143
Average Precision,433,0.010852451641925326
Average Precision,434,0.002551020408163265
Average Precision,435,0.02262963461382038
Average Precision,436,0.028251423059134598
Average Precision,438,0.01677592827690599
Average Precision,439,0.04701077174424722
Average Precision,440,0.17038995950286273
Average Precision,441,0.6486111111111111
Average Precision,442,0.010270990970239717
Average Precision,443,0.12274342016822896
Average Precision,445,0.0
Average Precision,446,0.02130996448778139
Average Precision,448,0.0

Average Precision,449,0.0416666666666666664
Average Precision,450,0.049333767966270765
Precision @10,401,0.1
Precision @10,402,0.3
Precision @10,403,0.5
Precision @10,404,0.0
Precision @10,405,0.1
Precision @10,406,0.4
Precision @10,407,0.3
Precision @10,408,0.4
Precision @10,409,0.0
Precision @10,410,0.3
Precision @10,411,0.6
Precision @10,412,0.2
Precision @10,413,0.0
Precision @10,414,0.2
Precision @10,415,0.1
Precision @10,417,0.1
Precision @10,418,0.4
Precision @10,419,0.1
Precision @10,420,0.6
Precision @10,421,0.0
Precision @10,422,0.2
Precision @10,424,0.3
Precision @10,425,0.3
Precision @10,426,0.2
Precision @10,427,0.1
Precision @10,428,0.0
Precision @10,429,0.1
Precision @10,430,0.3
Precision @10,431,0.6
Precision @10,432,0.0
Precision @10,433,0.0
Precision @10,434,0.0
Precision @10,435,0.1
Precision @10,436,0.4
Precision @10,438,0.1
Precision @10,439,0.1
Precision @10,440,0.1
Precision @10,441,0.5

Precision @10,442,0.2
Precision @10,443,0.2
Precision @10,445,0.0
Precision @10,446,0.1
Precision @10,448,0.0
Precision @10,449,0.1
Precision @10,450,0.0
NDCG @10,401,0.06943122193677727
NDCG @10,402,0.349966777951421
NDCG @10,403,0.5766882048947065
NDCG @10,404,0.0
NDCG @10,405,0.06943122193677727
NDCG @10,406,0.5682963021961281
NDCG @10,407,0.39375843764607205
NDCG @10,408,0.5384313152574521
NDCG @10,409,0.0
NDCG @10,410,0.8048099750039491
NDCG @10,411,0.6870165078530993
NDCG @10,412,0.16815228646891087
NDCG @10,413,0.0
NDCG @10,414,0.2836929289153804
NDCG @10,415,0.24630238874073
NDCG @10,417,0.13886244387355454
NDCG @10,418,0.34445239307234
NDCG @10,419,0.39038004999210174
NDCG @10,420,0.6339753813071975
NDCG @10,421,0.0
NDCG @10,422,0.20248323207250624
NDCG @10,424,0.3222722491219547
NDCG @10,425,0.39639187290150935
NDCG @10,426,0.14465249243306438
NDCG @10,427,0.2200917662980802
NDCG @10,428,0.0
NDCG @10,429,0.39038004999210174
NDCG @10,430,0.5773584151532217
NDCG @10,431,0.4362115423097744
NDCG @10,432,0.0
NDCG @10,433,0.0
NDCG @10,434,0.0
NDCG @10,435,0.07336392209936006

NDCG @10,436,0.3858930373209064
NDCG @10,438,0.06943122193677727
NDCG @10,439,0.13886244387355454
NDCG @10,440,0.2200917662980802
NDCG @10,441,0.81383546042969
NDCG @10,442,0.16421958630632805
NDCG @10,443,0.2863459897524693
NDCG @10,445,0.0
NDCG @10,446,0.06625422345438903
NDCG @10,448,0.0
NDCG @10,449,0.12647135138382856
NDCG @10,450,0.0
NDCG @1000,401,0.3453179622677145
NDCG @1000,402,0.5644287394368591
NDCG @1000,403,0.8043327944774392
NDCG @1000,404,0.20677703780378767
NDCG @1000,405,0.12192609118967469
NDCG @1000,406,0.8213458149293232
NDCG @1000,407,0.46983966017884604
NDCG @1000,408,0.5043246623774288
NDCG @1000,409,0.255958024809815 5
NDCG @1000,410,0.8693954474736921
NDCG @1000,411,0.5706678667406713
NDCG @1000,412,0.47003365567540917
NDCG @1000,413,0.13264079256781564
NDCG @1000,414,0.2836929289153804
NDCG @1000,415,0.24630238874073
NDCG @1000,417,0.31202555621883715
NDCG @1000,418,0.31634725600759656
NDCG @1000,419,0.5371844324883699
NDCG @1000,420,0.8025593814675847
NDCG @1000,421,0.1413805659746911
NDCG @1000,422,0.2434019049341932
NDCG @1000,424,0.2896330330370289
NDCG @1000,425,0.627370571519922
NDCG @1000,426,0.16958503154759783
NDCG @1000,427,0.22931594445056044
NDCG @1000,428,0.13136868206191152
NDCG @1000,429,0.39038004999210174
NDCG @1000,430,0.693624600381306

NDCG @1000,431,0.45056320819115575
NDCG @1000,432,0.08685168454816748
NDCG @1000,433,0.13057954254544646
NDCG @1000,434,0.08044384993556625
NDCG @1000,435,0.20648883759119666
NDCG @1000,436,0.1652333830278222
NDCG @1000,438,0.1476227415757353
NDCG @1000,439,0.18164658358740726
NDCG @1000,440,0.44949866281895007
NDCG @1000,441,0.81383546042969
NDCG @1000,442,0.11173460213428242
NDCG @1000,443,0.3852658276924744
NDCG @1000,445,0.0
NDCG @1000,446,0.19294029313468747
NDCG @1000,448,0.0
NDCG @1000,449,0.12647135138382856
NDCG @1000,450,0.3780722023346104
TBG,401,0.59445317530718
TBG,402,1.8397894097312117
TBG,403,3.4282219022581204
TBG,404,0.23123412364285958
TBG,405,0.8577155983523271
TBG,406,2.535299442565797
TBG,407,2.0596885387274013
TBG,408,4.3694579079197595
TBG,409,0.29989928021294365
TBG,410,1.3729268048807286
TBG,411,2.7014952842280766
TBG,412,1.639854063759949
TBG,413,2.660876150333566E-4
TBG,414,0.6950101585229458
TBG,415,0.45466224320881293
TBG,417,1.1062188702698845
TBG,418,2.3969150182623418
TBG,419,0.6699494300991664
TBG,420,4.489032464278202
TBG,421,0.005359541770094262
TBG,422,1.8289781634678017
TBG,424,2.158547129926882
TBG,425,3.530513401138586

TBG,426,1.4909087617165029
TBG,427,0.821667316336592
TBG,428,0.1113106214130516
TBG,429,0.4764708553148599
TBG,430,1.3410960926881268
TBG,431,2.488903394451743
TBG,432,0.04830885188994885
TBG,433,0.09518370718094714
TBG,434,1.4636408938510428E-4
TBG,435,0.49728194288720645
TBG,436,2.136088998087518
TBG,438,0.5871868729531045
TBG,439,0.44878127437239157
TBG,440,1.1589814633116309
TBG,441,2.042086314516553
TBG,442,0.8783986874429928
TBG,443,0.8878148933784986
TBG,445,0.0
TBG,446,0.5274382599408841
TBG,448,0.0
TBG,449,0.4445312234168459
TBG,450,0.5616240312021071

For Student 12:
Console output:

```
Unable to parse results file


Process finished with exit code 0
```

Files updated/created:
Updated: hw3-5a-jhhchen.csv

```
student12,bad format,bad format,bad format,bad format,bad format,
```

Question 6
The best student run was Student1. The results of a student's t-test are available in the t-test.csv file.

Effectiveness Measure	Best Run Score	msmuckerAND score	Relative Percent Improvement	Student's t-test, two-side, paired, p-value
Mean AP	0.25	0.090	177.778%	1.550e-7
Mean P@10	0.282	0.124	127.416%	3.470e-6
Mean NDCG@10	0.371	0.211	75.829%	4.189e-4
Mean NDCG@1000	0.485	0.272	78.309%	1.600e-7
Mean TBG	2.037	0.758	168.734%	5.680e-7

For each effectiveness measure, the p-values are all extremely low and significantly less than 0.05. Thus, the difference between the best run and the msmuckerAND results are statistically significant.

To determine the topics when BooleanAND is as good or better than the best run, a simple comparison for greater than or equal was done in Excel between the two sets of data. The table below summarizes the topics for which msmuckerAND had a score greater than or equal to Student 1's scores, aside from the topics where they both had a score of 0. This comparison is available in the *t-test.csv* file.

Effectiveness Measure	Topic
Average Precision	**410**
Average Precision	**415**
Average Precision	**421**
Precision @10	**408**
Precision @10	**410**
Precision @10	**415**
Precision @10	**421**
Precision @10	**424**
Precision @10	**426**

Precision @10	428
NDCG@10	**402**
NDCG@10	**408**
NDCG@10	**410**
NDCG@10	**421**
NDCG@10	**424**
NDCG@10	**426**
NDCG@10	**427**
NDCG@1000	**402**
NDCG@1000	**410**
NDCG@1000	**426**
NDCG@1000	**427**
NDCG@1000	436
TBG	**415**
TBG	**421**

If a topic using BooleanAND is considered to be better than the best run when it has two or more effectiveness measures that are better, then the following topics are better than the best run: 402, 408, 410, 415, 421, 424, 426, 427. Thus, BooleanAND retrieval is worse than the best student run because Student 1's scores are statistically significantly better on average, and only 8/45 of the topics were considered to be scored better by BooleanAND.