MSCI 541 HW3 Report

Ramandeep Farmaha 20516974

Problem 1

Using precision at rank 10 is more beneficial than using average precision in the cases where there is a significantly large number of relevant documents for a query, and the user only cares about the first page of results. An example would be a textbook database: a query such as "Microeconomics: Tenth Edition" would return potentially hundreds of results, with only the top books being relevant to the user. In this case, it's better to use precision at rank 10 as an evaluation metric because the user only cares about the top results and is unlikely to browse the next pages. If A is an information retrieval system that has a higher precision rank at 10 but lower average precision than system B for a particular query, it is more beneficial for the user to use system A, because she will only look at the first page of results, which are more relevant from system A.

Problem 2

Advantage 1: NDCG is much more fine tuned than precision rank at 10. For example, if there are two retrieval systems A and B, which rank the top ten documents for a given query as: {1, 0, 1, 1, 1, 0, 0, 1, 0, 1} and {1, 1, 0, 1, 1, 1, 1, 0, 0, 0} respectively, the precision ranks at 10 for both systems are identical, as they both contain 6 relevant documents in the top ten results. However, system B will have a higher NDCG than system A because its relevant results are clustered at the top of the list, making it a better option than system A. Web search engines would prefer to use NDCG, as it promotes models where the most relevant documents are clustered to the top.

Advantage 2: Precision rank at 10 assumes a binary representation of relevance: either the document is relevant or it isn't. NDCG, on the other hand, assigns a relevance score that varies in levels (perfect, excellent, good, fair, bad), which allows for a higher resolution of decision making. This is beneficial for a web search engine where a user may wish to access webpages that are tangentially related to their query (i.e. if a user queries for "Cairo" and a document pertaining to "Giza" is returned).

Problem 3

Part A

The randomization and bootstrap tests are appropriate to determine the statistical significance of the difference in the medians.

Part B

A non-paired test should be used, as 50 human participants are used for a study on system A and 50 *different* human participants are used for the same study on system B.

Part C

In order to conduct the experiment as a paired test, the same 50 human participants would need to be used to conduct the same study on both systems A and B.

Part D

For a large p-value of 0.8, we say "we failed to reject the null hypothesis". The difference between "we failed to reject the null hypothesis" and "we are forced to accept the null hypothesis" is that we are trying to prove an alternative hypothesis when we conduct a study, while the null hypothesis usually plays the role of the devil's advocate (i.e. no difference has been made). The study serves to provide evidence in proving the claim of the alternative hypothesis; and thus gives no evidence for the claim of the null hypothesis. Instead, we are forced to reject the alternative hypothesis but also don't have enough conclusive evidence to claim no difference has been made (i.e. accept the null hypothesis), thus we say "we failed to reject the null hypothesis".

Problem 4

Part A

A p-value of 0.06 means that for the 1000 samples, if Algorithms A and B were identical, there would be a 6 percent likelihood that that the observed mean difference (of 0.39-0.21 = 0.18) would occur. This translates to 60 sample mean differences for A and B being 0,18 in a set of 1000 samples.

Part B

I would recommend to conduct a randomization test to measure the statistical significance of the mean difference of B and C. If the test returns a p-value greater than 0.05, then I can safely reject algorithm B, as it produced a high p-value for both mean differences for A and C, and recommend using Algorithm C. However, if the test returns a p-value less than 0.05, I would recommend using algorithm B, becuase although it returned a p-value greater than the threshold for statistical significance for the mean difference between A and B, the p-value of 0.06 would be considered on the margin. The difference in the size of the mean difference between A and B vs A and C (0.18 vs 0.01) is large enough to warrant accepting a p-value of 0.06.

Problem 5

Part 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**	**2.033**
student2	0.141	0.193	0.251	0.344	1.250
student3	0.099	0.158	0.181	0.312	1.256
student4	0.202	0.244	0.328	0.427	1.756

Run Name	Mean Average Precision	Mean P@10	Mean NDCG@10	Mean NDCG@1000	Mean TBG
student5	*0.224*	0.256	0.320	*0.464*	*1.977*
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	1.866
student9	0.139	0.204	0.241	0.327	1.588
student10	bad format	bad format	bad format	bad format	bad format
student11	0.137	0.167	0.210	0.299	1.128
student12	bad format	bad format	bad format	bad format	bad format
student13	0.073	0.093	0.095	0.201	0.768
student14	0.200	0.251	0.323	0.415	1.756
msmuckerAND	0.098	0.133	0.161	0.273	0.819

Part C

Effectiveness Measure	Best Run Score	Second Best Run Score	Relative Percent Improvement	Two-sided Paired t-Test p-value
Mean Average Precision	0.250	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%	0.193
Mean TBG	2.033	1.977	2.833%	0.551

Part D

None of the p-values in the table in Part C were below 0.05, thus no * indications.

Part E

Console Input:

```
python3 evaluate.py --qrel qrels/LA-only.trec8-401.450.minus416-423-437-444-447.txt
--results_files results-files/student2.results results-files/student12.results
--output_directory=files_output/
```

Console Output:

```
Computing measures for: student2
Computing measures for: student12
Cannot print: student12, bad format
Summary stats for Mean Average Precision
Only a single row, possibly due to bad format
Summary stats for Mean P@10
Only a single row, possibly due to bad format
Summary stats for Mean NDCG@10
Only a single row, possibly due to bad format
Summary stats for Mean NDCG@1000
Only a single row, possibly due to bad format
```

```
Summary stats for Mean TBG
Only a single row, possibly due to bad format
```

The command generated 3 files in the `files_output` directory: `student2.csv`, `average_measures.csv`, and `summary_statistics.csv`. The file `student12.csv` was not created, since the `student12.results` file contains poorly formatted results, and thus cannot be analyzed.

`student2.csv`:

```
measure,query_id,score
precision_at_10,401,0.1
precision_at_10,402,0.3
precision_at_10,403,0.5
precision_at_10,404,0.0
precision_at_10,405,0.1
precision_at_10,406,0.4
precision_at_10,407,0.3
precision_at_10,408,0.4
precision_at_10,409,0.0
precision_at_10,410,0.3
precision_at_10,411,0.6
precision_at_10,412,0.2
precision_at_10,413,0.0
precision_at_10,414,0.2
precision_at_10,415,0.1
precision_at_10,417,0.1
precision_at_10,418,0.4
precision_at_10,419,0.1
precision_at_10,420,0.6
precision_at_10,421,0.0
precision_at_10,422,0.2
precision_at_10,424,0.3
precision_at_10,425,0.3
precision_at_10,426,0.2
precision_at_10,427,0.1
precision_at_10,428,0.0
precision_at_10,429,0.1
precision_at_10,430,0.3
precision_at_10,431,0.6
precision_at_10,432,0.0
precision_at_10,433,0.0
precision_at_10,434,0.0
precision_at_10,435,0.1
precision_at_10,436,0.4
precision_at_10,438,0.1
precision_at_10,439,0.1
precision_at_10,440,0.1
precision_at_10,441,0.5
precision_at_10,442,0.2
precision_at_10,443,0.2
precision_at_10,445,0.0
precision_at_10,446,0.1
precision_at_10,448,0.0
```

```
precision_at_10,449,0.1
precision_at_10,450,0.0
ndcg_10,401,0.06943122193677727
ndcg_10,402,0.3499667779514209
ndcg_10,403,0.5766882048947064
ndcg_10,404,0.0
ndcg_10,405,0.06943122193677727
ndcg_10,406,0.5682963021961281
ndcg_10,407,0.39375843764607193
ndcg_10,408,0.5384313152574521
ndcg_10,409,0.0
ndcg_10,410,0.8048099750039491
ndcg_10,411,0.6870165078530993
ndcg_10,412,0.1681522864689108
ndcg_10,413,0.0
ndcg_10,414,0.2836929289153804
ndcg_10,415,0.24630238874073
ndcg_10,417,0.13886244387355454
ndcg_10,418,0.34445239307233994
ndcg_10,419,0.3903800499921017
ndcg_10,420,0.6339753813071974
ndcg_10,421,0.0
ndcg_10,422,0.20248323207250624
ndcg_10,424,0.3222722491219547
ndcg_10,425,0.3963918729015093
ndcg_10,426,0.14465249243306436
ndcg_10,427,0.22009176629808017
ndcg_10,428,0.0
ndcg_10,429,0.3903800499921017
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_10,435,0.07336392209936005
ndcg_10,436,0.38589303732090635
ndcg_10,438,0.06943122193677727
ndcg_10,439,0.13886244387355454
ndcg_10,440,0.22009176629808017
ndcg_10,441,0.81383546042969
ndcg_10,442,0.16421958630632802
ndcg_10,443,0.2863459897524692
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
average_precision,401,0.0403377583185201
average_precision,402,0.155595467805563
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,408,0.17613258578682506
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,428,0.011111111111111111
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,449,0.041666666666666664
average_precision,450,0.049333767966270765
ndcg_1000,401,0.34531796226771455
ndcg_1000,402,0.5644811891434851
ndcg_1000,403,0.8043327944774391
ndcg_1000,404,0.20677703780378764
ndcg_1000,405,0.12192609118967468
ndcg_1000,406,0.8213458149293233
ndcg_1000,407,0.46983966017884593
ndcg_1000,408,0.5043200013417638
ndcg_1000,409,0.2559580248098155
ndcg_1000,410,0.8693954474736921
ndcg_1000,411,0.5706678667406713
ndcg_1000,412,0.47003365567540917
ndcg_1000,413,0.13264079256781566
ndcg_1000,414,0.2836929289153804
ndcg_1000,415,0.24630238874073
ndcg_1000,417,0.3120255562188371
ndcg_1000,418,0.31634725600759656
ndcg_1000,419,0.5371844324883698
ndcg_1000,420,0.8025593814675845
ndcg_1000,421,0.14138056597469106
ndcg_1000,422,0.24340190493419323
ndcg_1000,424,0.28963303303702886
```

```
ndcg_1000,425,0.6273705715199219
ndcg_1000,426,0.16958503154759783
ndcg_1000,427,0.22931594445056047
ndcg_1000,428,0.131368682061915
ndcg_1000,429,0.3903800499921017
ndcg_1000,430,0.6936246003813059
ndcg_1000,431,0.45056320819115575
ndcg_1000,432,0.08685168454816747
ndcg_1000,433,0.13057954254544643
ndcg_1000,434,0.08044384993556623
ndcg_1000,435,0.20648883759119663
ndcg_1000,436,0.1652333830278222
ndcg_1000,438,0.1476227415757353
ndcg_1000,439,0.18164658358740723
ndcg_1000,440,0.4494986628189501
ndcg_1000,441,0.81383546042969
ndcg_1000,442,0.11173460213428242
ndcg_1000,443,0.38526582769247436
ndcg_1000,445,0.0
ndcg_1000,446,0.19294029313468744
ndcg_1000,448,0.0
ndcg_1000,449,0.12647135138382856
ndcg_1000,450,0.3780722023346104
time_based_gain,401,0.5944531752409108
time_based_gain,402,1.840325195777382
time_based_gain,403,3.42822190225812
time_based_gain,404,0.23123412361679388
time_based_gain,405,0.857715598352327l
time_based_gain,406,2.529538016962721
time_based_gain,407,2.0596862783938517
time_based_gain,408,4.369457850341307
time_based_gain,409,0.29989928021294365
time_based_gain,410,1.3729268048807286
time_based_gain,411,2.7014948959581644
time_based_gain,412,1.6388367425174093
time_based_gain,413,0.0002600645498906938
time_based_gain,414,0.6950101585229458
time_based_gain,415,0.45466224320881293
time_based_gain,417,1.1062188701586049
time_based_gain,418,2.391937794315837
time_based_gain,419,0.6699494300991664
time_based_gain,420,4.4881836613753086
time_based_gain,421,0.005359210301665796
time_based_gain,422,1.828603691684225
time_based_gain,424,2.15854712992662
time_based_gain,425,3.530495836645362
time_based_gain,426,1.4859933146215427
time_based_gain,427,0.8216673163363521
time_based_gain,428,0.11069093414676752
time_based_gain,429,0.4764708553148599
time_based_gain,430,1.341096092688127
time_based_gain,431,2.488560214765563
time_based_gain,432,0.04830881326322973
time_based_gain,433,0.09517688649226813
time_based_gain,434,0.00014078519695097028
time_based_gain,435,0.4847604905978662
time_based_gain,436,2.127290843567231
```

```
time_based_gain,438,0.587186575830487
time_based_gain,439,0.44878118434344116
time_based_gain,440,1.1588384326303098
time_based_gain,441,2.042086314516553
time_based_gain,442,0.8783986394669994
time_based_gain,443,0.8875007334391881
time_based_gain,445,0
time_based_gain,446,0.5274382598395211
time_based_gain,448,0
time_based_gain,449,0.4445312234168459
time_based_gain,450,0.5614938964259921
```

`average_measures.csv`:

```
Run Name,Mean Average Precision,Mean P@10,Mean NDCG@10,Mean NDCG@1000,Mean TBG
student2,0.141,0.193,0.251,0.344,1.250
student12,bad format,bad format,bad format,bad format,bad format
```

`summary_statistics.csv`:

```
Effectiveness Measure,Best Run Score,Second Best Run Score,Relative Percent
Improvement,Two-sided Paired t-Test p-value
Mean Average Precision,0.141,N/A,N/A,N/A
Mean P@10,0.193,N/A,N/A,N/A
Mean NDCG@10,0.251,N/A,N/A,N/A
Mean NDCG@1000,0.344,N/A,N/A,N/A
Mean TBG,1.250,N/A,N/A,N/A
```

Problem 6

The best student run across the board (using the table in Q5 A/B) is student1. The msmuckerAND results are compared to the student1 results using the two-sided paired t-test in the table below:

Effectiveness Measure	Best Run Score	Second Best Run Score	Relative Percent Improvement	Two-sided Paired t-Test p-value
Mean Average Precision	0.250	0.098	155.102%	4.96165542091736e-08
Mean P@10	0.282	0.133	112.030%	1.042575407799581e-05
Mean NDCG@10	0.371	0.161	130.435%	0.0
Mean NDCG@1000	0.485	0.273	77.656%	6.018278798735953e-08
Mean TBG	2.033	0.819	148.230%	1.7146520327264363e-06

The incredibly low p-values indicate that the differences are stastically significant: i.e. the student1 rankings perform much better than msmuckerAND.

The following table indicates which topics and measures the msmuckerAND run performed better than the student1 run:

measure	query_id	mssmuckerAND	student1
ndcg_1000	403	0.7756184235923638	0.7407152541425661
ndcg_10	405	0.09478836436955078	0.07336392209936005
ndcg_1000	407	0.6131471927654584	0.5205333955935466
ndcg_10	426	0.3282410109263901	0.07839826897867533
precision_at_10	426	0.4	0.1
ndcg_1000	426	0.2701962884746054	0.17668676350093626
ndcg_1000	427	1.0	0.381328792137582
ndcg_1000	436	0.4512376968469067	0.3095580221883342
ndcg_10	442	0.2588097889896128	0.06943122193677727
precision_at_10	442	0.3	0.1

From the table above, only topics 426 and 442 performed better at multiple performance measures for msmuckerAND. Thus, the quality of the Boolean And retrieval is significantly lower than the best student run, as Student1 performed better than msmuckerAND for almost all of the topics across almost all measures, save for a small subsete in the table above.