Thomas Enns - 20823674
MSCI 541
2022-10-27

HW2 – Report

1. Precision at rank 10 is a measure of the proportion of relevant documents within the top ten documents returned by your search engine. Average precision can be considered to be the area under the interpolated curve when you plot the precision (y) and recall (x) on a cartesian plot. Therefore, average precision is recall sensitive.
   One would want to use precision at rank 10 instead of average precision in a situation where recall is less important than precision. This could be, for example, searching up the definition of a word on google. When doing this, you want to ideally only consult one website to fulfil your information need. It is important for the top results to be exactly what you are looking for. It is much less important for all the web pages that define your word (hundreds of online dictionaries), to be returned. In this situation, you can find exactly what you want with terrible recall, meaning a very low average precision.
   This average precision would be lower than one that you would want when searching for a washing machine on an e-commerce website, where you would want to see most likely everything they have on offer and would feel cheated if you were not shown all the washing machines. In the second case, you would prefer average precision over precision at rank 10.

2. NDCG is normalized, discounted, cumulative, gain. As defined in question 1, precision at rank 10 is the proportion of relevant documents within the top ten documents returned by the search engine.
   The notion of a proportion of a “relevant” or “non-relevant” document used by precision at rank ten implies that gain must be binary. This is the first advantage of nDCG at rank 10; it takes into account graded relevance. This is a more realistic way to model search results as a document can be somewhat relevant to a user.
   The second advantage of nDCG at rank 10 is that it has a measure of discount. Users do not all reach the same rank and will not receive any gain from results past where they stop reading. Precision at rank 10 unrealistically assumes that users will look at all the first ten results and will count gain from results that would not be visited. NDCG is therefore a more accurate model of user behavior.

3.
   a. Randomization and Bootstrapping tests are appropriate to determine the statistical significance of the two medians because they can be used with any statistic (mean, median, mode) and are distribution free. A student’s T test only applies to means and thus cannot be applied to determine the statistical significance of the difference between two medians. The sign test can incorrectly predict significance and should not be used in IR research (Smucker, Alan, & Carterette, 2007).
   b. You should use a non-paired test of statistical significance because these are two independent groups doing the evaluations of the two systems. If it was the same group doing both evaluations then a paired test would be appropriate.

c. As mentioned in B, the same group would need to complete both evaluations for a paired T test to be appropriate. Meaning that one group of 50 users would evaluate both system A and system B. In this situation, to account for the effect of users having seen the evaluation questions already when evaluating their second search engine, the order in which they evaluate them would need to be different for two halves of the group, split randomly.

d. In a situation where we calculate a p-value of 0.8, we say that we “fail to reject the null hypothesis”. A p-value of 0.8 means that according to our statistical test (which can be flawed), there is an 80% chance that the difference in our results could have occurred randomly under the null hypothesis. This does not mean that you have proven the null hypothesis. The statement “we are forced to accept the null hypothesis” would only be appropriate if you had statistically significant data showing that the null hypothesis is true. In this case we only fail to have statistically significant data showing that it is untrue. Thus, we “fail to reject the null hypothesis”.

4.

a. The null hypothesis of the randomization test is that both systems are in fact the same and any variation in performance is caused by randomness. To evaluate this, the randomization test assigns random signs, to the difference between the nDCG of the two systems in question for each of the 50 topics and calculates the mean difference in nDCG. If 1000 samples are used, this is done 1000 different times with 1000 randomizations of the differences in nDCG for the 50 topics.
Thus, the randomization test found that the absolute value of the mean difference in nDCG for systems A and B is greater than |0.21-0.39| 6% of the time (60 out of 1000 times). The second test found that the absolute value of the mean difference in nDCG for systems A and C is greater than |0.21-0.22| 0.2% of the time (2 out of 1000 times).

b. In this situation, C is clearly statistically significant, beyond much doubt. However, it offers only a very small improvement in normalized discounted cumulated gain compared to A. This means that C is probably consistently offering a very small improvement. B on the other hand, offers a larger average improvement but is not quite statistically significant. This indicates a lot of variance in the differences in nDCG between A and B.
In this situation, I would recommend further testing. One measure of search engine quality used with one set of topics and one statistical test is hardly definitive. I would recommend using other measures such as MAP and perhaps different topic sets as well if they are available. This would give a clearer picture of what should be implemented.

5.

a. Table from hw3-5a-WatIAMUserID.csv (also contains bolding and italicization from “b”):

Run Name	Mean Average Precision	Mean P@10	Mean NDCG@10	Mean NDCG@1000	Mean TBG
msmuckerAND	0.098	0.133	0.17	0.202	0.818
student1	**0.25**	*0.282*	**0.371**	**0.485**	**2.03**
student2	0.141	0.193	0.251	0.344	1.25
student3	0.099	0.158	0.181	0.312	1.252
student4	0.202	0.244	0.328	0.427	1.753
student5	*0.224*	0.256	0.32	*0.464*	*1.976*
student6	bad format	bad format	bad format	bad format	bad format
student7	0.107	0.153	0.189	0.284	1.209
student8	0.213	0.26	*0.346*	0.438	1.864
student9	0.139	0.204	0.241	0.327	1.586
student10	bad format	bad format	bad format	bad format	bad format
student11	0.137	0.167	0.21	0.299	1.125
student12	bad format	bad format	bad format	bad format	bad format
student13	0.073	**0.747**	0.115	0.199	0.767
student14	0.2	0.251	0.323	0.414	1.753

b. See bolding and italics in table above.

c. See Table in D

d. See Table Below.

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	11.46%	0.171
Mean P@10	0.747	0.282	164.57%	0.001*
Mean NDCG@10	0.371	0.346	7.44%	0.248
Mean NDCG@1000	0.485	0.464	4.67%	0.193
MEAN TBG	2.03	1.976	2.722	0.569

e. The run for student 2 is below

```
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw3-thomaspaenns> python Evaluate.py LA-only.trec8-401.450.minus416-423-437-444-447.txt s
tudent2.results
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw3-thomaspaenns>
```

This produces two output files, one CSV file for computing purposes and one human readable TXT file.

student2-measures	2022-11-04 9:42 PM	Microsoft Excel Com...	4 KB
student2-measures	2022-11-04 9:42 PM	Text Document	8 KB

The human readable text file output is similar to the sample output provided for the assignment. Abbreviations are used for each metric: AP for average precision, P@10 for precision at rank 10, nDCG@10 and nDCG@1000 for normalised discounted cumulated gain at ranks 10 and 1000 respectively, and TBG for time biased gain. The full contents of the file can be found in Appendix A.

The run for student 12 is below:

```
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw3-thomaspaenns> python Evaluate.py LA-only.trec8-401.450.minus416-423-437-444-447.txt s
tudent12.results
Use correctly formatted results file please
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw3-thomaspaenns>
```

For this run, the results file is incorrectly formatted so the program simply prompts the user to use a correctly formatted results file.

6. This answer is written assuming that the instructions meant we should compare the best overall student run with the msmuckerAND run, not the individual best for each metric. The best student run was by far the run for Student 1. It had the top metric in 4 of the five categories and in the one where it was not the top, it was the second best. The metric chose was the student's two side, paired T-test. This is appropriate as the metrics being compared are means. The metrics for the student1 run and the msmuckerAND run are compared in the table below.

Effectiveness Measure	student1 Score	smuckerAND score	Relative Percent Improvement	Student's t-test, two-side, paired, p-value
Mean AP	0.25	0.098	155.42	0
Mean P@10	0.282	0.133	111.667	0
Mean NDCG@10	0.371	0.17	117.871	0
Mean NDCG@1000	0.485	0.202	140.396	0
MEAN TBG	2.03	0.818	148.017	0

Note that the values in the above table are rounded to three decimal places. As such, the P-values are all shown to be zero. Running my script without rounding showed that all the P values are around E^-6 order, meaning that there are non-zero values beginning at 6 decimal places. This means that the results are most definitely statistically significant. The large relative percentages of improvement indicate that the conventional search engine far surpasses a BooleanAnd search in terms of all these metrics. Overall, the BooleanAnd search had far too many searches with no results or no relevant results to measure up against the conventional algorithm.
There are very few topics where BooleanAnd performs better thanor equal to Student1. In some cases they are equal for some metrics. In the table below, you can see the delta between a metric for student1 and booleanAnd. In the green cells, it is above or equal to zero:

	MAP	P@10	Ndcg@10	NDCG@1000	TBG
401	-0.09339	-0.2	-0.16141	-0.39337194	-1.23495
402	-0.17095	-0.2	-0.20098	-0.48183076	-1.695
403	-0.02393	-0.2	-0.06753	0.03490317	-0.24765
404	-0.0104	0	0	-0.171972	-0.01044
405	-0.01215	0	0.021424	-0.08161878	-0.71378
406	-0.29029	-0.1	-0.28551	-0.29156091	-1.04361
407	-0.1322	-0.3	-0.28351	-0.40892291	-2.00739
408	-0.12036	-0.3	-0.35971	-0.37214776	-2.96821
409	-0.1	-0.1	-0.28906	-0.28906483	-0.35688
410	0	0	0	0	-0.01411
411	-0.17542	-0.3	-0.4441	-0.47233631	-1.97519
412	-0.36476	-0.6	-0.55275	-0.30135434	-5.11812
413	-0.08333	0	0	-0.27023815	-0.31937
414	-0.10539	-0.1	-0.20211	-0.34365427	-0.55178
415	-0.21429	0	-0.26025	-0.26025337	-0.26797
417	-0.28272	-0.6	-0.70077	-0.29626608	-2.83295
418	-0.26042	-0.6	-0.72733	-0.61805722	-4.72893
419	-0.325	-0.2	-0.35915	-0.35914753	-0.66207
420	-0.25076	-0.3	-0.22784	-0.28173609	-1.34795
421	-0.0168	0	0	-0.24663103	-0.25443
422	-0.34997	-0.5	-0.38785	-0.62950483	-7.15262
424	-0.14128	-0.1	-0.06362	-0.494983	-2.07012
425	-0.18883	0	-0.16592	-0.35314712	-1.29274
426	-0.00069	0.3	0.249843	-0.04705346	-0.22524
427	-0.05267	-0.1	-0.08514	-0.25152072	-1.14797
428	-0.06501	0	-0.05613	-0.19246116	-0.22102
429	-0.54861	-0.3	-0.53195	-0.53194603	-1.22056
430	-0.47359	-0.2	-0.36005	-0.44913488	-1.11408
431	-0.23476	-0.2	-0.21821	-0.45789924	-1.65665
432	-0.00167	0	0	-0.06787393	-0.05126
433	-0.00484	0	0	-0.10910973	-0.01684
434	-0.5157	-0.1	-0.61315	-0.60253691	-0.44725
435	-0.03831	0	0	-0.27550377	-0.55348
436	-0.05905	-0.4	-0.35546	-0.20662764	-2.25209
438	-0.08859	-0.1	-0.0784	-0.25990853	-1.18161
439	-0.0146	0	0	-0.1679477	-0.33279
440	-0.51778	-0.5	-0.62741	-0.47082615	-2.60215
441	-0.15884	-0.3	-0.22961	-0.00044724	-0.42876
442	-0.00871	0.2	0.189379	-0.126234	-0.32027
443	-0.05858	-0.1	-0.08852	-0.23714757	-0.6496
445	-0.21944	-0.2	-0.41618	-0.23787745	-0.6947
446	-0.02142	0	0	-0.19127231	-0.16258

448	-0.00925	0	0	-0.19489335	-0.02431
449	-0.00698	0	0	-0.0898461	-0.22572
450	-0.02717	0	-0.09985	-0.20690536	-0.11221

From this table, we can see that topic 410 is the only place where the algorithms seem to be on an equal footing. In some other topics, BooleanAnd is equal or better only for the metrics at rank 10, most likely because it may return far fewer documents.
Overall however, the Student1 run is far better by all metrics considered and this is statistically significant with p-values less than 0.001.

Appendix A – Full Output for Student 2

AP	401	0.0403377583185201
AP	402	0.155595467805563
AP	403	0.5181658314928408
AP	404	0.026792114695340503
AP	405	0.023218294051627383
AP	406	0.5396358524344804
AP	407	0.12691027321387646
AP	408	0.17613258578682506
AP	409	0.07142857142857142
AP	410	0.7028846153846153
AP	411	0.2835203570626717
AP	412	0.0969455240957383
AP	413	0.005405405405405406
AP	414	0.10833333333333334
AP	415	0.125
AP	417	0.05884848769805482
AP	418	0.07067448933608388
AP	419	0.28407014979905004
AP	420	0.48257872961484244
AP	421	0.005804936852296678
AP	422	0.03867659574599077
AP	424	0.05506923888302862
AP	425	0.2720934005508434
AP	426	0.018594560268947468
AP	427	0.05366500273711303
AP	428	0.011111111111111111
AP	429	0.25
AP	430	0.3990972950304047
AP	431	0.1421563816876285
AP	432	0.0026239052263011143
AP	433	0.010852451641925326
AP	434	0.002551020408163265
AP	435	0.02262963461382038
AP	436	0.028251423059134598
AP	438	0.01677592827690599
AP	439	0.04701077174424722
AP	440	0.17038995950286273
AP	441	0.6486111111111111
AP	442	0.010270990970239717
AP	443	0.12274342016822896
AP	445	0.0
AP	446	0.02130996448778139
AP	448	0.0
AP	449	0.041666666666666664
AP	450	0.049333767966270765
P@10	401	0.1
P@10	402	0.3
P@10	403	0.5
P@10	404	0.0
P@10	405	0.1
P@10	406	0.4
P@10	407	0.3
P@10	408	0.4
P@10	409	0.0
P@10	410	0.3
P@10	411	0.6
P@10	412	0.2
P@10	413	0.0
P@10	414	0.2
P@10	415	0.1
P@10	417	0.1
P@10	418	0.4
P@10	419	0.1
P@10	420	0.6
P@10	421	0.0
P@10	422	0.2
P@10	424	0.3
P@10	425	0.3
P@10	426	0.2
P@10	427	0.1
P@10	428	0.0
P@10	429	0.1
P@10	430	0.3
P@10	431	0.6
P@10	432	0.0
P@10	433	0.0
P@10	434	0.0
P@10	435	0.1
P@10	436	0.4
P@10	438	0.1
P@10	439	0.1
P@10	440	0.1
P@10	441	0.5
P@10	442	0.2
P@10	443	0.2
P@10	445	0.0
P@10	446	0.1
P@10	448	0.0
P@10	449	0.1
P@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
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	0.1313686820619115
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
TBG	401	0.5944531751747446
TBG	402	1.8403250255332442
TBG	403	3.42822190225812
TBG	404	0.2312341235930742
TBG	405	0.8577155983523271
TBG	406	2.5237021311587293
TBG	407	2.0596840601130197
TBG	408	4.369457841344936
TBG	409	0.29989928021294365
TBG	410	1.3729268048807286
TBG	411	2.7014945041411504
TBG	412	1.6378641018146731
TBG	413	0.0002543214186581846
TBG	414	0.6950101585229458
TBG	415	0.45466224320881293
TBG	417	1.1062188700376006
TBG	418	2.387557960279907
TBG	419	0.6699494300991664
TBG	420	4.4873838268175215
TBG	421	0.005358914219573441
TBG	422	1.828279653416333
TBG	424	2.158547129926396
TBG	425	3.530478129416083
TBG	426	1.48108953398607
TBG	427	0.821667316336073
TBG	428	0.11010817761988546
TBG	429	0.4764708553148599
TBG	430	1.341096092688127
TBG	431	2.4882102758727025
TBG	432	0.04830877756307518
TBG	433	0.09517058908482569
TBG	434	0.00013608540466778765
TBG	435	0.47363943347187254
TBG	436	2.1169648163008623
TBG	438	0.5871862856087394
TBG	439	0.4487811030201035
TBG	440	1.1587069633647902
TBG	441	2.042086314516553
TBG	442	0.8783985932395991
TBG	443	0.887183199253599
TBG	445	0.0
TBG	446	0.5274382597487242
TBG	448	0.0
TBG	449	0.4445312234168459
TBG	450	0.5613802038434605