MSCI 541 – Search Engines

Homework 3

Professor Mark Smucker

Matthew Erxleben

ID: 20889980

Date: November 7th, 2023

Problem 1)

Average precision is unsuitable for evaluation of web searches where the document collections are typically in billion or more documents. This is because average precision is recall-sensitive. The documents that are relevant that are not included in the ranked list, contribute zero to the summation for average precision. Average precision is calculated by:

AP = 1 / (Total relevant documents) * Summation from 1 to # of Results (binary relevance * precision at rank i)

Average precision requires us to know how many total relevant documents there are in the entire document collection. This is impossible for modern web searching, where only the user can know if the document is relevant or not, and there are billions of options. It is not realistic for the user to have seen every single document and evaluate if it is relevant or not.

Problem 2)

nDCG@10 and Precision at rank 10 both have their advantages and disadvantages. In terms of comparing the two as evaluation measures for a web search engine, here are two advantages nDCG@10 has over Precision at rank 10:

1. The first advantage nDCG@10 has over precision at rank 10 is that precision doesn't tell us if the search engine returns the items at a higher rank versus a lower rank. For example, if there are 2 relevant results for a query that the search engine returned in 10 results, the precision at rank 10 will be 2/10 = 0.2. However, this does not tell us if the search engine returned the relevant documents 1^{st} and 2^{nd}, or 9^{th} and 10^{th}. nDCG does tell us if the relevant documents were returned first, as the gain at rank i (the numerator of the DCG summation calculation) will be greater than 0 if the document has gain towards the user, and is 0 otherwise. Therefore, in IDCG, the most relevant documents come first. This is compared to the DCG of our search engine, using the nDCG ratio to see how close the search engine is to returning the documents with the most gain first.
2. The second advantage nDCG@10 has over precision at rank 10 is that nDCG can represent variable relevance, whereas precision can only represent binary relevance. This is because nDCG utilizes gain at rank, and the gain the user can get from a document is not restricted to just 0 or 1. However with precision, for the calculation, a document can either be relevant or not. This relies on binary relevance to be able to be calculated, therefore there is no middle ground on if a document is somewhat relevant.

Problem 3)

3a)

I would use student t-tests to determine the statistical significance of the difference in the medians.

3b)

This problem would require a non-paired test of statistical significance.

3c)

In 3b, I answered that the problem would require a non-paired test of statistical significance. To utilize a paired test, some changes would need to be made to the experiment. A paired test requires pairs of data. One of the main issues with this experiment using a paired test is the fact that none of the participants who used system A also used system B. In a paired test, it would be required that the participants utilize both system A and system B and now we can test the statistical significance of their results against each other for each participant's results in the different systems. This is a way to change the experiment, which will now allow us to use a paired test.

3d)

If we obtain a large p-value of 0.8, we should say "We failed to reject the null hypothesis". This is because the large p-value is likely greater than our alpha (typically 0.05), and therefore the difference between the two systems is not statistically significant. Here is an example of why we say we fail to reject the null hypothesis. Let's say we have two Olympic sprinters; Usain Bolt and Andre De Grasse. Furthermore, we have collected data on each of their performances for how fast they ran the same races, at the same time, in the same place. The hypothesis is that "These two runners are the same speed, therefore the same person". If we find that one of the racers on average is faster, and utilize a t-test to see that the difference is statistically significant, then we can **reject** that they are the same speed and the same person. This is rejecting the null hypothesis. However, if we cannot prove that one racer is statistically significantly faster than the other, then we cannot assume they are the same speed and the same person. This does not mean that they are the same person, but it also does not mean that they are not the same person. Therefore, we **fail to reject the null hypothesis.** If we were to say "We are forced to accept the null hypothesis", we would be saying that because we could not prove one of the racers is statistically significantly faster than the other, Usain Bolt and Andre De Grasse must be the same speed and the same person. This is not true; therefore, we say the first option instead.

Problem 4)

Ranking Algorithm	Mean nDCG
A	0.21
B	0.39
C	0.22

Experiment	Mean nDCG Difference	Relative Percent Improvement	p-value
A vs B	+0.18	85.71%	0.06
A vs C	+0.01	4.76%	0.002

4a)

When doing a randomized test with 1000 samples to measure the statistical significance of the mean nDCG difference between algorithms A and B, the p-value is 0.06. This means that in those 1000 samples, 6% of them had a mean nDCG difference that is the same or even larger than the observed mean nDCG difference. In this example, our null hypothesis would be that algorithms A and B are **equal** in terms of ranking algorithms, and the alternative hypothesis is that B is the better algorithm. Due to our p-value being higher than 0.05, this means that this test is not statistically significant. This means that we cannot reject the null hypothesis, as we do not have strong evidence to prove that the two algorithms are not equal. Although we see the large difference in Mean nDCG is likely showing us that there is a large difference between the two algorithms, this is not statistically significant due to the p-value being so large. Therefore, this does not mean the two algorithms are not equal, but it also does not mean that they are equal.

4b)

As we know, the p-value shows that the mean difference between algorithms A and C is statistically significant. This means we can reject the null hypothesis that the two algorithms are equal, as the difference between the two is statistically significant. Therefore, there is a slight improvement from algorithm A to C. The change in mean nDCG is very small, however, that is still a significant difference that can be made to immediately improve the ranking algorithm. At companies like Google, they are making tons of changes to the ranking algorithm every single day. These are small and incremental changes to the algorithm that are statistically significant. Although this is a startup, and a more significant improvement is important when it comes to our algorithm. Even if an improvement is statistically significant, the first question that we should ask ourselves is “is this change a significant improvement?”. If the improvement is so small, it does not matter if it is statistically significant or not because the user will not be affected by the change and anyway. Therefore, I would choose to start to use ranking algorithm C, but would be on the fence between just staying with algorithm A due to the difference being not very

significant. Especially if the overhead to implement algorithm C is a lot. In the future, I would also recommend using different experiments to test the difference between algorithms. For example, testing the precision at a certain rank. This can be another way to evaluate your search engine (worse than nDCG, but still another evaluation tactic). The difference between algorithms A and B is quite large (85.71% Relative Percent Improvement), which likely is an outlier. This can be seen as the p-value is so high, therefore showing this improvement to not be statistically significant. In the future, I would recommend resampling and observing the mean nDCG for algorithm B. Then comparing algorithm B's performance to algorithm C, would be an effective measure to see if the changes are statistically significant and if there is an improvement with algorithm B.

Problem 5)

Installation Requirements:

1. Please make sure Python is installed on your computer before running the program.

2. Clone the repository on your device by entering this into your terminal: git clone https://github.com/UWaterloo-MSCI-541/msci-541-f23-hw3-matterxleben.git

Downloading the hw3 files:

Navigate to the LEARN page and download the hw3-files-2023.zip file. Please unzip this file to a location on your desktop.

Running the Programs:

To run these programs, please navigate to where you cloned the repository and open the working directory .../msci-541-f23-hw3-matterxleben

ResultsEvaluation.py:

The program accepts 2 command line arguments: the directory location of your results file, the directory location of your qrels file:

For example, you would run ResultsEvaluation.py from the command prompt / terminal / shell as:

python ResultsEvaluation.py C:/Users/matth/OneDrive/Desktop/University/3B/MSCI541-Search-Engines/HW3/hw3-files-2023/results-files/student12.results

C:/Users/matth/OneDrive/Desktop/University/3B/MSCI541-Search-Engines/HW3/hw3-files-2023/qrels/LA-only.trec8-401.450.minus416-423-437-444-447.txt

5a)

Run	Mean AP	Mean P@10	Mean NDCG@10	Mean NDCG@1000
student1	**0.250**	**0.282**	**0.371**	**0.485**
student2	0.141	0.193	0.251	0.344
student3	0.099	0.158	0.181	0.312
student4	0.202	0.244	0.328	0.427
student5	*0.224*	0.256	0.320	*0.464*
student6	BAD FORMAT	BAD FORMAT	BAD FORMAT	BAD FORMAT
student7	BAD FORMAT	BAD FORMAT	BAD FORMAT	BAD FORMAT
student8	0.213	*0.260*	*0.346*	0.438
student9	0.139	0.204	0.241	0.327
student10	BAD FORMAT	BAD FORMAT	BAD FORMAT	BAD FORMAT
student11	0.137	0.167	0.210	0.299
student12	BAD FORMAT	BAD FORMAT	BAD FORMAT	BAD FORMAT
student13	0.073	0.093	0.093	0.199
student14	0.200	0.251	0.323	0.414
msmuckerAND	0.098	0.133	0.104	0.202

As we can see, student 6, 7, 10, and 12 are BAD FORMAT. This is because they are not in the proper format, therefore we cannot analyze the results. For example, student7.results had a donco that was not in the proper format, therefore my code outputted this to the terminal and ended the program:

Student7.results

```
PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben>  c:; cd 'c:\Users\matth\OneDrive\Desktop\Uni
versity\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben'; & 'C:\Users\matth\AppData\Local\Programs\Python\Python39\python.exe' 'c:\Users\ma
tth\.vscode\extensions\ms-python.python-2023.18.0\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launcher' '59465' '--' 'C:\Users\matth\OneDrive\
Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben\ResultsEvaluation.py'
Topic: 401 AP: 0.0108 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.1817
Topic: 402 AP: 0.0516 P@10: 0.2 NDCG@10: 0.2489 NDCG@1000: 0.224
Topic: 403 AP: 0.3603 P@10: 0.2 NDCG@10: 0.359 NDCG@1000: 0.7234
Topic: 404 AP: 0.0 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.0

                    This docno: LA082189-002 is not properly formatted! The docno must be 13 characters long in the format: LA123456-7890

                    Therefore we must cancel the results for this file and exit the program!

PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben>
```

Student 6, 10, and 12 results:

```
PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben>  c:; cd 'c:\Users\matth\OneDrive\Desktop\Uni
versity\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben'; & 'C:\Users\matth\AppData\Local\Programs\Python\Python39\python.exe' 'c:\Users\ma
tth\.vscode\extensions\ms-python.python-2023.18.0\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launcher' '63295' '--' 'C:\Users\matth\OneDrive\
Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben\ResultsEvaluation.py'

                    This docno: [LA021890-010is not properly formatted! The docno must be 13 characters long in the format: LA123456-7890

                    Therefore we must cancel the results for this file and exit the program!

PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben>  c:; cd 'c:\Users\matth\OneDrive\Desktop\Uni
versity\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben'; & 'C:\Users\matth\AppData\Local\Programs\Python\Python39\python.exe' 'c:\Users\ma
                    Therefore we must cancel the results for this file and exit the program!

PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben>  c:; cd 'c:\Users\matth\OneDrive\Desktop\Uni
versity\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben'; & 'C:\Users\matth\AppData\Local\Programs\Python\Python39\python.exe' 'c:\Users\ma
tth\.vscode\extensions\ms-python.python-2023.18.0\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launcher' '63306' '--' 'C:\Users\matth\OneDrive\
Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben\ResultsEvaluation.py'

                    This docno: la101790-0075is not properly formatted! The docno must be 13 characters long in the format: LA123456-7890

                    Therefore we must cancel the results for this file and exit the program!

PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben>  c:; cd 'c:\Users\matth\OneDrive\Desktop\Uni
versity\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben'; & 'C:\Users\matth\AppData\Local\Programs\Python\Python39\python.exe' 'c:\Users\ma
tth\.vscode\extensions\ms-python.python-2023.18.0\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launcher' '63316' '--' 'C:\Users\matth\OneDrive\
Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben\ResultsEvaluation.py'
This line: topic, Q0 0, rank, score, student12
is not properly formatted! Each line must be in the format:
                    401 Q0 LA021890-0100 1 1 username
                    This line does not have a numeric topic number in the first 3 characters.
                    Therefore we must cancel the results for this file and exit the program!

PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben>
```

5b)

The highest score is bolded and the second highest score is italicized in the table.

5c)

Measure	Best run score	Second best run score	Relative Percent Improvement	Student's t-test, two-side, paired, p-value
Mean AP	0.250	0.224	11.46%	0.170662
Mean P@10	0.282	0.260	8.55%	0.242866
Mean NDCG@10	0.371	0.346	7.43%	0.248466
Mean NDCG@1000	0.485	0.464	4.67%	0.193475

5d)

As we can see, the p-values are computed for each effectiveness measure. This shows the difference between the two retrievals is not statistically significant, because $p > 0.05$. This shows that the student1.results are not statistically significant, therefore none of the scores have the * symbol.

5e)

student2.results:

```
PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben> c
versity\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben'; & 'C:\Users\matth\AppData\Local\Program
tth\.vscode\extensions\ms-python.python-2023.18.0\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launch
Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben\ResultsEvaluation.py'
Topic: 401 AP: 0.0403 P@10: 0.1 NDCG@10: 0.0694 NDCG@1000: 0.3453
Topic: 402 AP: 0.1556 P@10: 0.3 NDCG@10: 0.35 NDCG@1000: 0.5645
Topic: 403 AP: 0.5182 P@10: 0.5 NDCG@10: 0.5767 NDCG@1000: 0.8043
Topic: 404 AP: 0.0268 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.2068
Topic: 405 AP: 0.0232 P@10: 0.1 NDCG@10: 0.0694 NDCG@1000: 0.1219
Topic: 406 AP: 0.5396 P@10: 0.4 NDCG@10: 0.5683 NDCG@1000: 0.8213
Topic: 407 AP: 0.1269 P@10: 0.3 NDCG@10: 0.3938 NDCG@1000: 0.4698
Topic: 408 AP: 0.1761 P@10: 0.4 NDCG@10: 0.5384 NDCG@1000: 0.5043
Topic: 409 AP: 0.0714 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.256
Topic: 410 AP: 0.7029 P@10: 0.3 NDCG@10: 0.8048 NDCG@1000: 0.8694
Topic: 411 AP: 0.2835 P@10: 0.6 NDCG@10: 0.687 NDCG@1000: 0.5707
Topic: 412 AP: 0.0969 P@10: 0.2 NDCG@10: 0.1682 NDCG@1000: 0.47
Topic: 413 AP: 0.0054 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.1326
Topic: 414 AP: 0.1083 P@10: 0.2 NDCG@10: 0.2837 NDCG@1000: 0.2837
Topic: 415 AP: 0.125 P@10: 0.1 NDCG@10: 0.2463 NDCG@1000: 0.2463
Topic: 417 AP: 0.0588 P@10: 0.1 NDCG@10: 0.1389 NDCG@1000: 0.312
Topic: 418 AP: 0.0707 P@10: 0.4 NDCG@10: 0.3445 NDCG@1000: 0.3163
Topic: 419 AP: 0.2841 P@10: 0.1 NDCG@10: 0.3904 NDCG@1000: 0.5372
Topic: 420 AP: 0.4826 P@10: 0.6 NDCG@10: 0.634 NDCG@1000: 0.8026
Topic: 421 AP: 0.0058 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.1414
Topic: 422 AP: 0.0387 P@10: 0.2 NDCG@10: 0.2025 NDCG@1000: 0.2434
Topic: 424 AP: 0.0551 P@10: 0.3 NDCG@10: 0.3223 NDCG@1000: 0.2896
Topic: 425 AP: 0.2721 P@10: 0.3 NDCG@10: 0.3964 NDCG@1000: 0.6274
Topic: 426 AP: 0.0186 P@10: 0.2 NDCG@10: 0.1447 NDCG@1000: 0.1696
Topic: 427 AP: 0.0537 P@10: 0.1 NDCG@10: 0.2201 NDCG@1000: 0.2293
Topic: 428 AP: 0.0111 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.1314
Topic: 429 AP: 0.25 P@10: 0.1 NDCG@10: 0.3904 NDCG@1000: 0.3904
Topic: 430 AP: 0.3991 P@10: 0.3 NDCG@10: 0.5774 NDCG@1000: 0.6936
Topic: 431 AP: 0.1422 P@10: 0.6 NDCG@10: 0.4362 NDCG@1000: 0.4506
Topic: 432 AP: 0.0026 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.0869
Topic: 433 AP: 0.0109 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.1306
Topic: 434 AP: 0.0026 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.0804
Topic: 435 AP: 0.0226 P@10: 0.1 NDCG@10: 0.0734 NDCG@1000: 0.2065
Topic: 436 AP: 0.0283 P@10: 0.4 NDCG@10: 0.3859 NDCG@1000: 0.1652
Topic: 438 AP: 0.0168 P@10: 0.1 NDCG@10: 0.0694 NDCG@1000: 0.1476
Topic: 439 AP: 0.047 P@10: 0.1 NDCG@10: 0.1389 NDCG@1000: 0.1816
Topic: 440 AP: 0.1704 P@10: 0.1 NDCG@10: 0.2201 NDCG@1000: 0.4495
Topic: 441 AP: 0.6486 P@10: 0.5 NDCG@10: 0.8138 NDCG@1000: 0.8138
Topic: 442 AP: 0.0103 P@10: 0.2 NDCG@10: 0.1642 NDCG@1000: 0.1117
Topic: 443 AP: 0.1227 P@10: 0.2 NDCG@10: 0.2863 NDCG@1000: 0.3853
Topic: 445 AP: 0.0 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.0
Topic: 446 AP: 0.0213 P@10: 0.1 NDCG@10: 0.0663 NDCG@1000: 0.1929
Topic: 448 AP: 0.0 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.0
Topic: 449 AP: 0.0417 P@10: 0.1 NDCG@10: 0.1265 NDCG@1000: 0.1265
Topic: 450 AP: 0.0493 P@10: 0.0 NDCG@10: 0.0 NDCG@1000: 0.3781
PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben>
```

As we can see, the output is in the format:

Topic: ___ AP: ___ P@10: ___ NDCG@10: ___ NDCG@1000: ___

student12.results:

student12.results file is not in the proper formatting as we can see here:

```
topic, Q0 0, rank, score, student12
401,la021890-0100,0,1, Q0 student12
 student12
401,la122990-0070,0,2, Q0 student12
 student12
401,la082690-0052,0,3, Q0 student12
 student12
401,la040490-0003,0,4, Q0 student12
 student12
401,la050590-0114,0,5, Q0 student12
 student12
401,la051390-0170,0,6, Q0 student12
 student12
401,la040389-0047,0,7, Q0 student12
 student12
401,la052190-0065,0,8, Q0 student12
 student12
401,la111289-0073,0,9, Q0 student12
 student12
401,la100889-0019,0,10, Q0 student12
 student12
401,la090490-0093,0,11, Q0 student12
 student12
401,la050789-0068,0,12, Q0 student12
 student12
```

Therefore, the program does not output results to the terminal. Instead, the program reports that the file is incorrectly formatted. This bad run is noted in the report. This run is not allowed to be analyzed due to its incorrect formatting.

```
PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben>  c:
versity\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben'; & 'C:\Users\matth\AppData\Local\Programs
tth\.vscode\extensions\ms-python.python-2023.18.0\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launche
Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben\ResultsEvaluation.py'
This line: topic, Q0 0, rank, score, student12
is not properly formatted! Each line must be in the format:
                    401 Q0 LA021890-0100 1 1 username
                    This line does not have a numeric topic number in the first 3 characters.
                    Therefore we must cancel the results for this file and exit the program!

PS C:\Users\matth\OneDrive\Desktop\University\3B\MSCI541-Search-Engines\HW3\msci-541-f23-hw3-matterxleben>
```

Problem 6)

The best student for all effectiveness measures is student 1. Student 1 is compared to msmuckerAND in the table below. As we can see, I utilized the t-test to compare the statistical significance of the difference between these two results. As we can see, the p-values computed are extremely small for each effectiveness measure. This shows the difference between the two retrievals is statistically significant ($p < 0.05$), showing student1.results to be much better than msmuckerAND.results.

Measure	Best run score	msmuckerAND score	Relative Percent Improvement	Student's t-test, two-side, paired, p-value
Mean AP	0.250	0.098	155.41%	0.00000004966540061
Mean P@10	0.282	0.133	111.67%	0.00001042575408
Mean NDCG@10	0.371	0.104	257.54%	0.000000258881056
Mean NDCG@1000	0.485	0.202	140.40%	0

To see if there are some topics for which Boolean AND is as good or better than student1's run, I compared each topic's mean effectiveness measure. I utilized Excel to highlight if the two are equivalent, then the column will equal 1 and highlight yellow. If msmuckerAND BooleanAND retrieval is better than student 1's run, then the column equals 2 and highlights green.

Student1					msmuckerAND					Comparison			
Topic	Mean AP	Mean P@10	Mean NDCG@1	Mean NDCG@1	Topic	Mean AP	Mean P@10	Mean NDCG@1	Mean NDCG@1	Mean AP	Mean P@10	Mean NDCG@1	Mean NDCG@1000
401	0.1039	0.3	0.2466	0.4502	401	0.0105	0.1	0.0851	0.0568	0	0	0	0
402	0.2086	0.4	0.4499	0.6026	402	0.0376	0.2	0	0.1208	0	0	0	0
403	0.5075	0.6	0.5302	0.7407	403	0.4836	0.4	0.4627	0.7756	0	0	0	2
404	0.0104	0	0	0.172	404	0	0	0	0	0	1	1	0
405	0.026	0.1	0.0734	0.1471	405	0.0139	0.1	0.0948	0.0654	0	1	2	0
406	0.4408	0.3	0.4575	0.7648	406	0.1505	0.2	0.172	0.4732	0	0	0	0
407	0.1667	0.4	0.5036	0.5205	407	0.0345	0.1	0	0.1116	0	0	0	0
408	0.1362	0.3	0.3597	0.4625	408	0.0158	0	0	0.0903	0	0	0	0
409	0.1	0.1	0.2891	0.2891	409	0	0	0	0	0	0	0	0
410	1	0.4	1	1	410	1	0.4	0	1	1	1	0	1
411	0.1754	0.3	0.4441	0.4723	411	0	0	0	0	0	0	0	0
412	0.4674	0.8	0.7163	0.7355	412	0.1026	0.2	0.1635	0.4341	0	0	0	0
413	0.0833	0	0	0.2702	413	0	0	0	0	0	1	1	0
414	0.1054	0.1	0.2021	0.3437	414	0	0	0	0	0	0	0	0
415	0.25	0.1	0.3904	0.3904	415	0.0357	0.1	0	0.1301	0	1	0	0
417	0.355	0.7	0.7792	0.7321	417	0.0723	0.1	0.0784	0.4359	0	0	0	0
418	0.2604	0.6	0.7273	0.6181	418	0	0	0	0	0	0	0	0
419	0.575	0.3	0.7495	0.7495	419	0.25	0.1	0	0.3904	0	0	0	0
420	0.6197	0.8	0.8604	0.8873	420	0.369	0.5	0.6325	0.6056	0	0	0	0
421	0.0188	0	0	0.2854	421	0.002	0	0	0.0388	0	1	1	0
422	0.3525	0.6	0.4826	0.6504	422	0.0026	0.1	0	0.0209	0	0	0	0
424	0.1535	0.1	0.0636	0.5924	424	0.0123	0	0	0.0974	0	0	0	0
425	0.4829	0.7	0.7624	0.8244	425	0.294	0.7	0.5965	0.4713	0	1	0	0
426	0.0342	0.1	0.0784	0.1767	426	0.0335	0.4	0.3282	0.1296	0	2	2	0
427	0.0961	0.2	0.3052	0.3813	427	0.0435	0.1	0	0.1298	0	0	0	0
428	0.1067	0.1	0.1952	0.3315	428	0.0417	0.1	0.1391	0.1391	0	1	0	0
429	0.7986	0.4	0.9223	0.9223	429	0.25	0.1	0	0.3904	0	0	0	0
430	0.6203	0.4	0.6608	0.7499	430	0.1467	0.2	0	0.3008	0	0	0	0
431	0.3165	0.6	0.6064	0.6818	431	0.0817	0.4	0.3882	0.2239	0	0	0	0
432	0.0017	0	0	0.0679	432	0	0	0	0	0	1	1	0
433	0.0048	0	0	0.1091	433	0	0	0	0	0	1	1	0
434	0.5435	0.1	0.6131	0.7469	434	0.0278	0	0	0.1443	0	0	0	0
435	0.0383	0	0	0.2755	435	0	0	0	0	0	1	1	0
436	0.0857	0.7	0.6777	0.3096	436	0.0267	0.3	0.3223	0.1029	0	0	0	0
438	0.1104	0.1	0.0784	0.4503	438	0.0218	0	0	0.1903	0	0	0	0
439	0.0146	0	0	0.1679	439	0	0	0	0	0	1	1	0
440	0.5684	0.5	0.6274	0.8132	440	0.0507	0	0	0.3424	0	0	0	0
441	0.6496	0.6	0.7634	0.7634	441	0.4908	0.3	0.5338	0.763	0	0	0	0
442	0.0229	0.1	0.0694	0.1964	442	0.0142	0.3	0.2588	0.0702	0	2	2	0
443	0.1034	0.2	0.1737	0.4177	443	0.0448	0.1	0.0851	0.1805	0	0	0	0
445	0.2444	0.2	0.4162	0.4162	445	0.025	0	0	0.1783	0	0	0	0
446	0.0253	0	0	0.2338	446	0.0038	0	0	0.0425	0	1	1	0
448	0.0092	0	0	0.1949	448	0	0	0	0	0	1	1	0
449	0.007	0	0	0.0898	449	0	0	0	0	0	1	1	0
450	0.2379	0.4	0.4315	0.6477	450	0.2108	0.4	0.3317	0.4408	0	1	0	0

As we can see, here are the results that are equal or greater for Boolean AND:

Mean AP: 410

Mean P@10: 404 405 410 413 415 421 425 426 428 432 433 435 439 442 446 448 449 450

Mean NDCG@10: 404 405 413 421 426 432 433 435 439 442 446 448 449

Mean NDCG@1000: 403 410

In conclusion, we can see the quality of Boolean AND is much worse than student1's run. Student1 uses modern retrieval methods, while Boolean AND is known for being a bad retrieval method, due to its binary retrieval nature.