MSCI 541 Homework 3

Thomas Kleinknecht
Tkleinkn, 20883814

Problem 1

Average precision is unsuitable for evaluation of web search where the document collections are a billion or more documents as average precision must be calculated with the total number of relevant documents in the collection, as it is recall sensitive. This is nearly impossible for a document collection of this size, leaving the extremely expensive and therefore redundant option of going through all documents, or a “best guess” at the total number of relevant documents, rendering this irrelevant.

Problem 2

1. Precision has no sensitivity to rank or order within the first 10, and nDCG @ 10 takes rank into account through discounting. Because of this, if documents 1-5 were not relevant, and 6-10 were relevant, precision would return 5/10, and the same with the opposite case. These two cases would be drastically different for nDCG @ 10.
2. nDCG allows us to quantify relevance, by assigning a gain which can take various values, such as 0,1, 2, 3 for example. Precision, however, is based on binary relevance and therefore does not allow for indication of varying levels of relevance between documents.

Problem 3

a) It is not suggested to use the sign test when studying results in information retrieval (Smucker, reference 1), and it is used with paired data, therefore we can immediately disregard this. The student’s t-test is only appropriate for testing differences between means, therefore it can be disregarded as well. Due to the fact that both the randomization and bootstrap tests can be used for any test statistic and neither requires paired data, either are appropriate to use.

b) You should use a non-paired test, as there are no common factors across pairs of data between systems, such as the same participant using each system, or each system performing the same task or search.

c) As I explained in b), in order to need a paired test, we would need a strong common factor across both systems, and this could be achieved by having each participant find an answer with both of the systems, so that we could compare the difference between the times of each system for each participant, giving us n pairs of data for n participants.

d) We say that we failed to reject the null hypothesis. This is because in this context, we can only reject the null hypothesis, but do not have the tools to confirm that it is true, or accept it. All that we can definitely say, is that given a certain test, we think it is unlikely (reasonably defined by $p <$ given value) that the null hypothesis is not true, we do not have the evidence at this time to reject this truth. This is because, given an individual test, we can disprove a hypothesis, however, proving it is true is nearly impossible and would require a whole host of complex testing. An example would be addition of numbers. If we test 1 + 1, and find that it does not equal 2, we can reject our null hypothesis of addition holding. If we find that it does equal 2, this does not prove that addition always holds, as what about all of the other cases, of 1+2, 4+5, etc., therefore we cannot definitely accept addition as the truth.

Problem 4

a) This p-value of 0.06, with respect to 1000 samples, for a randomization test, tells us that theoretically, in 60 of these 1000 samples, the difference between nDCG of A and nDCG of B should be greater or equal to 0.18. This could be that the value for A is over 0.18 units higher than B or vice versa. The relatively low occurrence of this tells us that the difference between the average of A and the average of B is very unlikely to have just happened by chance.
b) I would recommend that we change to algorithm B rather than C, or some other course of action. This is the correct decision due to two factors. First, how close the p-value is to being generally accepted as statistically significant (difference of 0.01), meaning that a very minimal risk is being taken. Second, the improvement in nDCG offered by algorithm B is substantial and nearly doubles the gain, or value, which is provided to the customer.

Problem 5

How to build and run code:

In order to build and run the program, the program will need to be cloned from the repository, and can be run in the command line/terminal if you have java downloaded on your computer (or after downloading).

First, ensure you have the proper directories and files set up. You will need the following arguments ready:

1. File path to directory which contains all of your files for this program
2. File path from the directory in 1. to the qrels file, including its name (without "/" before)
3. File path from the directory in 1. to the results file, including its name (without "/" before)
4. File path from the directory in 1 to the directory in which you would like to put your output (without "/" before)
5. Name of output file (.txt)

And you will also need to put the queries file in the clones repository into the directory in 1 above.

Running MeasureResults:

In order to run MeasureResults, navigate to the cloned repository, and compile the code using the command:

javac MeasureResults.java

Now you can run MeasureResults with following command:

java MeasureResults argument1 argument2 argument3 argument4 argument5

For example, locally I run it with this:

java MeasureResults "/Users/thomaskleinknecht/Desktop/MSCI 541/HW3" "hw3-files-2023/qrels/LA-only.trec8-401.450.minus416-423-437-444-447.txt" "hw3-files-2023/results-files/student7.results" "MeasureResultsOutput" "student7.measures.txt"

These arguments should be enclosed in quotations if there are any spaces in the directory or file names.

This program will output a file, named as you chose, to the output directory of your choice. In the first row of this file, it will show you the format of the following lines of the document. Each topic in the document is on its own line, with each of the effectiveness measure values listed next to that topic name, divided by a space between each.

In the event that the program runs into an issue with the result file's formatting, it will throw an error, output a message to the terminal, and output a message to the output file explaining that these results cannot be measured.

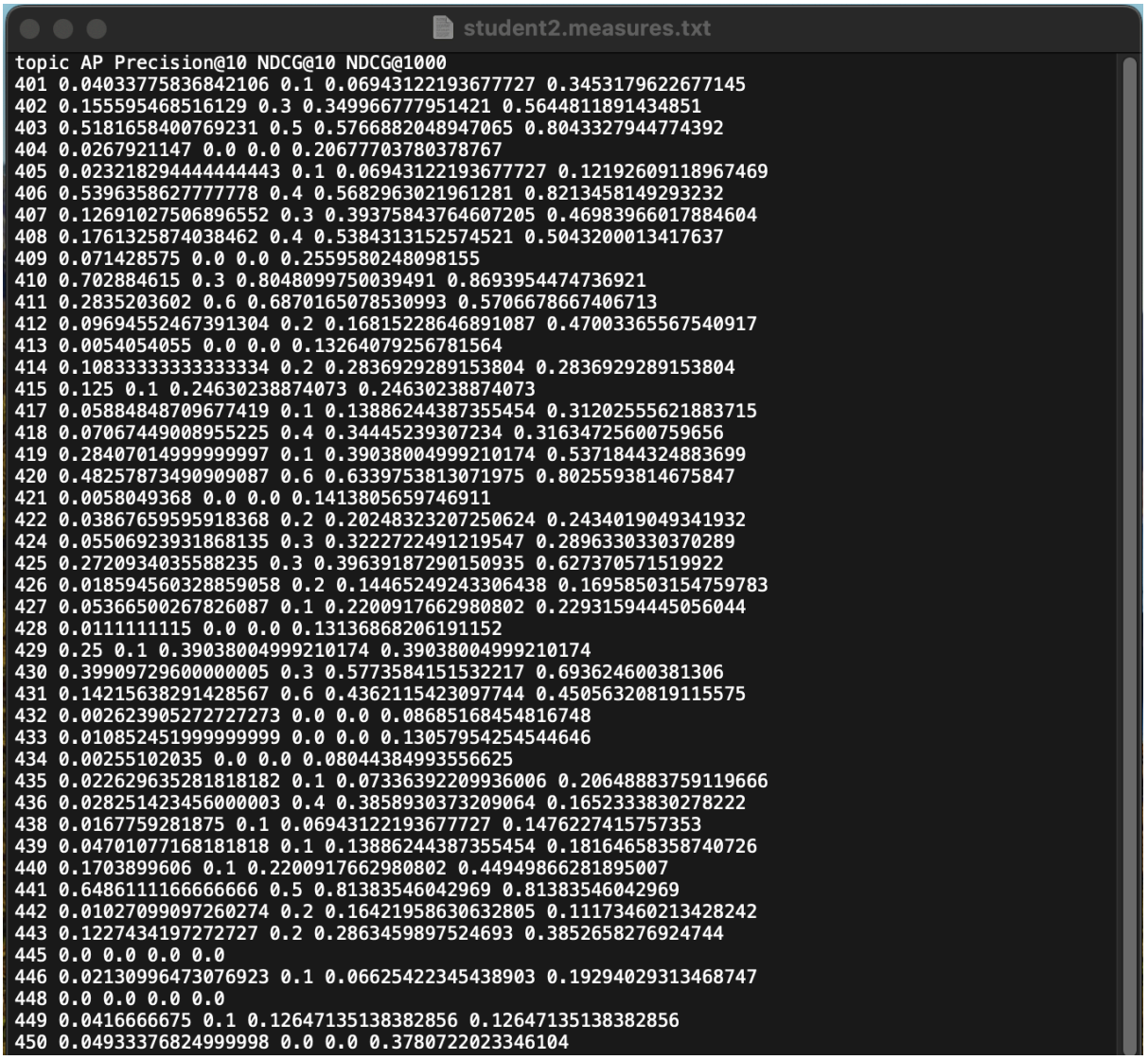
a) Average Results and b) Bolded and Italicized best and second best for each

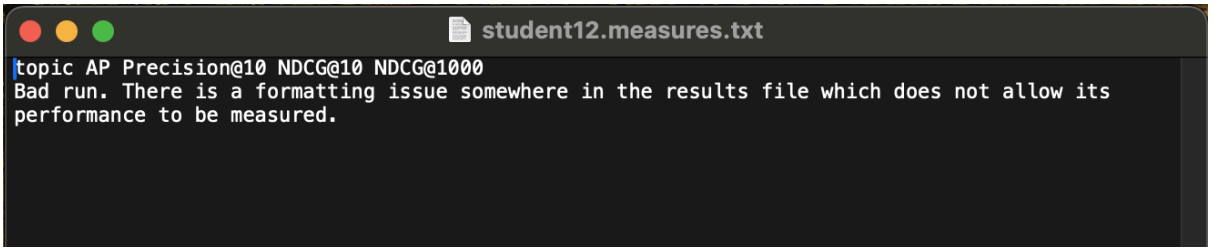
Run Name	Mean Average Precision	Mean P@10	Mean NDCG@10	Mean NDCG@1000
student1	**0.250**	**0.282**	**0.371**	**0.485**
student2	0.141	0.193	0.251	0.344
student3	0.099	0.158	0.181	0.312
student4	0.202	0.244	0.328	0.427
student5	*0.224*	0.256	0.320	*0.464*
student6	bad format	bad format	bad format	bad 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	0.115	0.199
student14	0.200	0.251	0.323	0.414
msmuckerAND	0.098	0.133	0.170	0.202

c) and d) Comparison of best and second best with student's t-test

Effectiveness Measure	Best Run Score	Second Best Run Score	Relative Percent Improvement	Student's t-test, two-sided, paired, p-value
Mean AP	0.25	0.224	11.61%	0.171
Mean P@10	0.282	0.26	8.46%	0.243
Mean NDCG@10	0.371	0.346	7.23%	0.248
Mean NDCG@1000	0.485	0.464	4.53%	0.193

e) Results of running student1.results and student12.results individually through program





Problem 6

Effectiveness Measure	Student 1 Score (Best Student)	msmuckerAND Score	Relative Percent Improvement	Student's t-test, two-sided, paired, p-value
Mean AP	0.25	0.098	155.10%	0.00000004961655
Mean P@10	0.282	0.133	112.03%	0.00001042575408
Mean NDCG@10	0.371	0.17	118.24%	0.00000057364357
Mean NDCG@1000	0.485	0.202	140.10%	0.00000000000001

I chose to use the student's t-test to compare paired data points for the measures for each of the search topics. The p-values can be seen in the right-most column in the above table. The differences for each of the effectiveness measures are statistically significant, as each of the p-values is < 0.05.

When comparing topic by topic, I found that Boolean AND does just as good or better on a few of the search topics in certain effectiveness measures:

- Topic 403: NDCG@1000 is just as good or better for Boolean AND
- Topic 405: Precision@10 and NDCG@10 are just as good or better for Boolean AND
- Topic 410: All four effectiveness measures are just as good or better
- Topic 415: Precision@10 is just as good or better for Boolean AND
- Topic 425: Precision@10 is just as good or better for Boolean AND
- Topic 426: Precision@10 and NDCG@10 are just as good or better for Boolean AND
- Topic 428: Precision@10 is just as good or better for Boolean AND
- Topic 442: Precision@10 and NDCG@10 are just as good or better for Boolean AND
- Topic 450: Precision@10 is just as good or better for Boolean AND

Overall, we can see that Boolean AND is much worse than the best student run. It does however, occasionally yield effectiveness measures which compare to the student run. Only 1 of the 45 search topics was equal or better across all four measures however, showing how even in the varying needs of all of the other queries, Boolean AND was worse at servicing these needs every time.

References

1. Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. https://doi.org/10.1145/1321440.1321528
2. Heavily referenced pseudocode from course content, as well as design suggestions made in campuswire.