Programming Assignment - Part A

Information Retrieval 1 [2018/2019]

Deadline: Wednesday, January 16th, 17:00
Submit: A PDF document with the design of the experiment in a clear step-by-step fashion, and how this design can answer our question. **Important Note**: The assignment asks for the experimental design ONLY and not about the implementation/results.
Filename to submit: <id1>_<id2>_<id3>_design.pdf

The homework covers the topics covered in Lecture 1 – 3.

Commercial search engines typically use a funnel approach in evaluating a new search algorithm: they first use an offline test collection to compare the production algorithm (P) with the new experimental algorithm (E); if *E* outperforms *P* with respect to the evaluation measure of their interest, the two algorithms are then compared online through an interleaving experiment.

In an interleaving experiment the ranked results of *P* and *E* (against a user query) are interleaved in a single ranked list which is presented to a user. The user then clicks on the results and the algorithm that receives most of the clicks wins the comparison. The experiment is repeated for a number of times (impressions) and the total wins for *P* and *E* are computed.

A Sign/Binomial Test is then run to examine whether the difference in wins between the two algorithms is statistically significant. Alternatively one can calculate the proportion of times *E* wins and test whether this proportion, p, is greater than $p_0=0.5$. This is called an 1-sample 1-sided proportion test.

The offline test collection is static (i.e. fixed), however when it comes to the online experiment, one would like to know for how long the experiment should run, or in other words how many impressions are necessary so that a statistically significant difference between *E* and *P* can be detected. E.g. if out of 100 impressions *E* wins 55% of them, can this still be due to random chance?

Unfortunately we cannot know a priori (i.e. before actually running the experiment) the proportion of *E* wins against *P.* If we knew we could perform a power analysis of the proportion

test and find the necessary number of impressions. What we know however is by what margin E outperformed P in the offline experiments.

For the purpose of this homework we assume that: (a) documents in the ranked list are being judged on a 3-graded scale, i.e. {0, 1, 2}, and (b) the evaluation measure of interest is the Expected Reciprocal Rank at cut-off rank 5 (ERR@5).

Further, for the purpose of this homework we assume that, (a) there are two interleaving algorithm we are interested in, the Team-Draft Interleaving, and the Probabilistic Interleaving, (b) users click on documents on the basis of the Position-Based Model.

In this homework we will determine the sample size (i.e. the number of interleaved impressions) required for a Team-Draft/Probabilistic Interleaving experiment to identify statistical significant differences, when it is known that in offline evaluation E outperformed P by a certain margin measured by the Expected Reciprocal Rank. In particular, we are looking at filling in the following tables.

ΔERR	Number of Impressions Team-Draft Interleaving		
	Min	Median	Max
[0.05, 0.1)			
[0.1, 0.2)			
[0.2, 0.3)			
...			
[0.8, 0.9)			
[0.9, 0.95]			

ΔERR	Number of Impressions Probabilistic Interleaving		
	Min	Median	Max
[0.05, 0.1)			
[0.1, 0.2)			
[0.2, 0.3)			
...			
[0.8, 0.9)			
[0.9, 0.95]			

Design and describe an experiment, step-by-step, which would allow filling in the tables above. Make your description as clear as possible, in a way that would allow a different team to implement and run your experiment. Make any assumptions you think necessary and describe them explicitly. Argue how the proposed experiment can answer the question of the assignment.

Important Note: The assignment asks **ONLY** for the design and **NOT** the implementation or results – you do not need to fill out the above tables, only describe an experiment to be used to fill them out. The assignment may sound to some extend "vague", since it does not specify "how to do things". This vagueness is by design, so you can come up with your own solutions.