Homework Assignment № 3

Prof. Joao Palotti (jpalotti@andrew.cmu.edu), CMU-Q Due to **1/May/2017**

1 Assignment Goals

This assignment aims to introduce you to how scientific research is done. Your two tasks are:

- Task 1: Get familiarized on how research with search technology is presented;
- Task 2: Create a scientific paper documenting your experiments.

2 Task 1 (10 points) – Individual Task

As stated and emphasized by Lecture 8 (see `https://github.com/joaopalotti/cmu_67300/blob/master/slides/lecture8_evaluation_campaigns.pdf`), the main body of research in search technologies was made through annual campaigns as TREC.

For the first part of this assignment, you will have to read a three research papers and summarize their findings. Your summary should contain a clear state on the task that they are trying to solve and their main approach, idea or hypothesis. Describe what kind of pre-processing decisions were taken, what retrieval methods were tested, and what conclusions were reached. Pick 3 out this list of papers:

- AT&T at TREC 9 (2000): `http://trec.nist.gov/pubs/trec9/papers/att-trec9.pdf`
- Microsoft Research Asia at Web TREC 2001: `http://trec.nist.gov/pubs/trec10/papers/msra.trec10.pdf`.
- WIDIT at TREC-2003 Web Track: `http://trec.nist.gov/pubs/trec12/papers/indianau.web.pdf`.
- Indri at TREC 2004: Terabyte Track: `http://trec.nist.gov/pubs/trec13/papers/umass.tera.pdf`.
- Dublin City University at the TREC 2005 Terabyte Track: `http://trec.nist.gov/pubs/trec14/images/pdf.gif`.
- RMIT University at TREC 2006: Terabyte Track: `http://trec.nist.gov/pubs/trec15/papers/rmit.tera.final.pdf`
- Lucene and Juru at TREC 2007: 1-Million Queries Track: `http://trec.nist.gov/pubs/trec16/papers/ibm-haifa.mq.final.pdf`
- A Study of Term Proximity and Document Weighting Normalization in Pseudo Relevance Feedback–UIUC at TREC 2009 Million Query Track: `http://trec.nist.gov/pubs/trec18/papers/uiuc.MQ.pdf`

- MMCI at the TREC 2010 Web Track: http://trec.nist.gov/pubs/trec19/papers/saarland.univ.web.rev.pdf
- Microsoft Research at TREC 2011 Web Track: http://trec.nist.gov/pubs/trec20/papers/msrsv.web.update.pdf (consider only the ad-hoc submissions, i.e. ignore the diversity task submissions)

3 Task 2 (10 points)

We finally have the assessments for the 18 queries that we created. They can be found at: https://github.com/joaopalotti/cmu_67300/blob/master/project_code/simple.qrels.

Your goal is to download these assessments and use the program named *trec_eval* (http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz) to evaluate your results. Provide a documentation such as the ones that you just read in Task 1 of this assignment. Use the -q parameter to collect information about each individual query. Your documentation should have:

- Author Names and Affiliation: name of the group participants and your email addresses.
- Abstract: simply state that your experiments are conducted for course CMU 67-300.
- Introduction: briefly introduce the task (Ad-hoc search) and the collection (numbers for our simpleWiki collection).
- Methods: describe your pre-processing steps, your implementation decisions and your runs.
- Result Analysis: describe your results in form of tables or graphs. Were all results equally good/bad? Did any method worked better than others? Can you tell why? Have you looked at the results of individual queries to find the weakness of your system?
- Conclusion: State your conclusion and describe what you would do better or try different if you had much more time (future work).

4 Submissions

Generate and send me one PDF file for <u>each</u> of the tasks above. The PDF files generated **have** to follow the ACM format, that can be obtained at https://www.acm.org/publications/proceedings-template. I highly recommend you to use the LaTeX version. Consider also using some online and collaborative environment such as:

- sharelatex.com (https://www.sharelatex.com/templates/552d98adeee6edb00c043d2f);
- overleaf.com (https://www.overleaf.com/latex/templates/association-for-computing-machinery-bmvfhcdnxfty#.WPR1nnX5g1L)