188.992 Experiment Design - WS 2020/21
Exercise 2:

Reproduce experimental results from a paper

Unsupervised Ranking - Entity Set Search of Scientific Literature

by Jiaming Shen, et al., Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA

Group 14:
Judith Lukács, 01125956
Tobias Hajszan, 11776172
Moritz Staudinger, 1777768

Key Findings & Encountered Difficulties

- Inconsistent folder naming in scripts: TREC-BIO / TREC_BIO
- Relative paths: wrong depth
- Encoding of script
- Multiple READMEs
- Missing necessary configuration for elasticsearch (necessary without explicit mentioning)
- SetRank claims to be significantly better, our first experiment does not support this claim
- Differentiation between TREC-BIO and TREC-BIO-ESQ not clear in repository

Paper Overview

- Experiments for enhancing quality of (scientific) literature search engines
- Comparison of efficiency of new algorithm on bag-of words/-entities mechanism
- Unsupervised Ranking algorithm on entity-sets
- Two standardized data sets (S2CS, TREC-BIO)
- Testing against baseline algorithms (BM25, LM-DIR, LM-JM, IB)
- Two-tailed t-test with p-value <= 0.05 (significance threshold)
- Further test with real life conditions for biomedical literature search based on PubMed queries

Strategy

- Read & Understand Paper → What does the researchers aim to prove?
- 1st run → try to reproduce paper/experiment one-to-one examine shortcomings of documentation
- Make necessary adjustments & reproduce experiment → comparison of output
- Diverging results → Identification of mistakes / What went wrong?

- 2nd run → try to reproduce experiment AGAIN with altered settings → accordance?
- If yes GREAT, if not → Why? Reasoning!
- Contact authors / report shortcomings / propose improvement / pull request to repository

Conclusions & Remaining Work

- So far reproducible with minor adjustments
  - path/working directory correction
  - file-name adaption
  - improved encoding
  - custom configuration of *elaticsearch*
- Intermediate results partly equal to paper results
- TODO:
  - identify shortcomings of current implementation
  - re-run experiment to rule out prior mistakes
  - comparison of latter results → if still no accordance documentation of ideas why/what went wrong
  - contacting authors