3.12 CodaLab

Evelyne Viegas (Microsoft Research – Redmond, US)

Main reference https://github.com/codalab

CodaLab is an open source platform which goal is to accelerate the rate of research by enabling collaboration among researchers and scientists across disciplines and make science truly reproducible. CodaLab Worksheets[13] focuses on accelerating data-driven research and making it more sound while enabling scientists to publish their research as executables papers with full provenance on data and code. CodaLab competitions[14] is a powerful framework for running data-driven competitions that involve result and/or code submission. Users can either participate in an existing competition or host a new competition as an organiser. CodaLab enables coopetitions, a new collaborative framework where users with different expertise can work together in a new environment favouring cross-pollination of ideas.

4 State of the Art in Different Areas of CS

4.1 State of the art trade-offs in IR Research

Shane Culpepper (RMIT University – Melbourne, AU)

URL https://github.com/lintool/IR-Reproducibility

This talk briefly presented a state-of-the-art comparison of ad-hoc search engines for a common TREC task. By aggregating results from the IR Reproducibility Challenge in the 2015 ACM SIGIR Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR), we contrast fully reproducible baseline runs and "best known" submissions from the TREC Adhoc Search Task between 2004–2006.

4.2 Managing and Curating IR Experimental Data

Nicola Ferro (University of Padova, IT)

Information Retrieval (IR) is a discipline deeply rooted in experimentation since its inception and, over the time, it has developed robust and shared methodologies for conducting experiments, relying on the so-called Cranfield Paradigm. In particular, the adoption of large-scale and shared experimental collections, typically used in international evaluation

[13] https://github.com/codalab/codalab-worksheets/wiki
[14] https://competitions.codalab.org/

campaigns like TREC[15], CLEF[16], and NTCIR[17] and then available for further re-use by the community, provide the means for running comparable experiments. This experimental paradigm gives raise to three targets for reproducibility:

- **experimental collections**: they consists of documents, topics, which surrogate real user information needs, and relevance judgements, which determine which documents are relevant to which topics. Experimental collections are an integral part of the experimental design and they are often used for many different purposes after their creation. It is thus important to understand their limitations and their generalizability as well as the reproduce the process that led to their creation. This is not always trivial since, for example, topics may be sampled from real system logs or relevance judgements are made by humans and, more and more often, using crowdsourcing.
- **system runs**: they are the most common target for reproducibility since they are what is discussed in papers proposing new methods and algorithms.
- **meta-evaluation experiments**: IR has a strong tradition in assessing its own evaluation methodologies, such as robustness of the experimental collections, reliability of the adopted evaluation measures or appropriateness of the adopted statistical analysis methods. All these investigations strongly rely on existing experimental collections and gathered systems runs and their reproducibility should be a key concern, since they probe our own experimental methods.

All the above mentioned three targets for reproducibility heavily depend on experimental data. Unfortunately, even if IR has a long tradition in ensuring the due scientific rigor is guaranteed in producing such data, it has not a similar tradition in managing and taking care of such valuable data. There currently are several barriers to proper data curation for reproducibility. There is a lack of common formats for modelling and describing the experimental data as well as almost no metadata (descriptive, administrative, copyright, etc.) for annotating and enriching them. The semantics of the data themselves is often not explicit and it is demanded to the scripts typically used for processing them, which are often not well documented, rely on rigid assumptions on the data format or even on side effects in processing the data. Finally, IR lacks a commonly agreed mechanism for citing and linking data to the papers describing them.

All these issues may be addressed by adapting solutions developed in other fields with similar problems but the biggest issue is the community itself, which would need to evolve its experimental methodologies to take into account reproducibility and the actions needed to guarantee it. This calls for an orchestrated effort and a cultural change which are the most compelling challenges towards a proper management and curation of experimental data.

[15] http://trec.nist.gov/
[16] http://www.clef-initiative.eu/
[17] http://research.nii.ac.jp/ntcir/index-en.html