Title: A Reproduction of Continuous Active Learning

Original Authors: Gordon V. Cormack and Maura R. Grossman. “Evaluation of machine-learning protocols for technology-assisted review in electronic discovery”, SIGIR 2014.

Reproduced Paper Authors: Zhe Yu, Nicholas A. Kraft, Tim Menzies. “Finding better active learners for faster literature reviews”, EMSE 2018.

What: Continuous active learning (CAL) is an algorithm proposed by Cormack and Grossman to solve the total recall problem in electronic discovery. CAL aims to help attorneys retrieving relevant documents (often only 1%) with least cost. Unlike conventional active learning algorithms, which utilize strategies such as uncertainty sampling to select training examples that can improves the trained model most, CAL focuses on retrieving examples that current model is most certain on. In this manner, CAL queries human oracles least but retrieve most relevant documents.

Why: Similar total recall problem exists in software engineering (SE) literature reviews. As SE researchers, much of our time is spent writing SE research papers. All that effort is wasted if no one find and read those papers. Studies show that median citations/year in SE research papers are disturbingly low: 0.23 to 0.33 citations per author in conference papers and 0.17 to 0.27 citations per author in journal papers. These are bad numbers since they show that most SE research papers are not found and not used by subsequent work. This is a terrible state of affairs since it explains all too well why industrial practitioners cannot find research that is relevant to their day-to-day work. As a result, helping SE researchers better and faster find relevant research papers become a crucial task in advancing and propagating SE researches. This is a similar task as electronic discovery. Usually, SE researchers screen thousands of titles and abstracts to include dozens of relevant studies. As certainly being one of the state of the art solution to total recall problem, CAL is being reproduced when Yu et al. tried to solve this SE finding relevant research papers problem.

How: The reproduction of CAL is based on the description in the original paper. While the original dataset is not freely available online, the reproduced algorithm is tested on SE literature review data. Yu et al. found that by mix and match CAL with two other state of the art active learning algorithm, they were able to create a new algorithm called FASTREAD that outperforms all the prior state of the art ones including CAL by 20-50%. FASTREAD is able to find 95% of the relevant SE papers by asking humans to review 5-30% of the candidate papers.

Where: Reproduced package of CAL along with the FASTREAD algorithm is available at
https://doi.org/10.5281/zenodo.837861 or https://github.com/fastread/src.

Discussion: It would have made the reproduction much easier If the original authors provided pseudocode and dataset for CAL. In future, we would like to recommend that every research should be open source including the code, data, along with the results. This will reduce time and effort to reproduce the code and verify the implementations. Also, having a pre-print version of the paper available would largely reduce the time of other researchers knowing, using, and reproducing the results. This will facilitate researches in SE and advance the field faster.