Gunja Sarkar

CSCI 317: People-Powered Procedures

Spring 2022

Project 1 Write-Up

The open-source software I chose to look at is OCR4all – an open-source software that performs optical character recognition specifically for historical texts. OCR4all described a complete OCR workflow, starting with pre-processing of the image files, to layout segmentation, line segmentation, text recognition, and corrections of the recognized text. In my experience, OCR4all is an accomplished project – it’s comprehensive and works well. However, in some ways, it seems “near completion.” This is to say there doesn’t seem to be much of an opportunity to directly improve the source code of the software. Even if there is the option to do so, there is a distinct lack of organization or clarity for a casual user to figure out how they can contribute besides simply providing more training cases and test cases. That being said, OCR4all should perhaps instead be viewed as a different kind of open-source project – one that is funded by an educational or governmental institution to completion and then left open and available to the use and discretion of the public.

OCR4all was developed by Julius-Maximilians-Universität Würzburg under the direction of Christian Reul as part of the Kallimachus joint project which is a joint project looking to intentionally bridge the divide between computer science and the humanities. For example, the board of directors for the Kallimachus project consists of a musicology professor, a philosophy professor, a computer philology and modern German literature professor, and two computer science professors. The project was approved in July of 2021 to be funded for two years by the Federal Ministry of Education to be continues and institutionalized in the Center for Philology and Digitality. The project was designed and created by Christian Reul who is still involved with the project and project management today. Most of the staff who have routinely worked and collaborated on this project seem to be other members of the academic community, mostly students and some former professors. There is no code of conduct, but I think that may be in part because OCR4all is seen primarily as a educational tool. OCR4all has always been licensed for open use for educational purposes and there are still no restrictions apparent on the GitHub repo and website. The project seems to have begun in 2020 and recent contributions to the source code has certainly slowed down after peaking in late 2020 and 2021. In the summer of 2020, OCR4all worked with OCR-D, an OCR project funded by the German Research Foundation which is an educational institution largely funded by the national German government. OCR-D has a more specific focus on digitizing 16th to 18th century German-speaking area texts, but the OCR-D ideas were successfully implemented and adapted in OCR4all for more extensive use.

The project itself is interesting to try to explore as an English speaker. While the software and much of the documentation was prepared initially in German, almost all components are accessible to English users. In fact, using OCR4all is completely feasible and possible as an English speaker. The project has a getting\_started GitHub repository with set-up guides and comprehensive step-by-step user manuals in English. If anything, it seems as though there is a little too much guidance because it’s hard to know where to start. The getting\_started repository also advertises a mailing list for users to join to stay informed about the developments and new version releases. I joined the mailing list upon my initial exploration of this project and still have not yet received anything or been given notice that I was approved to actually be added to the mailing list. The project itself is cumbersome to set up. It is necessary to install a virtual machine simulator to run the software (VirtualBox, by Oracle), Docker (a GitHub-esque file storage server), and then a starting OCR4all compressed file which was 3.3 GB and took almost an hour to download on my computer. Once you have all the needed components, it is definitely a little confusing to navigate the different parts of the project and figure out how to put the pieces of the project together, but it is certainly doable, even for perhaps not the most technical of users. The user guide is worthy of commendation – it’s quite extensive and includes sufficiently layman explanations of the different aspects of the project. If anything, as previously mentioned, the excess of instructions is possibly more intimidating and confusing. I think OCR4all would certainly benefit from a more streamlined starting process. Contributing to the actual source code is much more difficult. Because of the use of a virtual machine, the user’s role in the process is mostly to improve the source code with more test samples and model training to improve the accuracy of the model. I think this is largely in part because the project has reached a sufficient stage of completion where it no longer requires constant maintenance and has been institutionalized by a particular organization, it’s no longer “open-source” in the way something is crowdsourced, but rather open to the public for use and distribution. It certainly is a component of the project that is not quite clear until you’ve downloaded and installed gigabytes of setup software and sample data. Users can still and are encouraged to provide input by contacting the creators by email and the workflow and the model can be trained by users, but the model itself has reached a stage where there isn’t really work left to do. Thus, OCR4all is a good example of how the term “open-source” is now used in context different than perhaps originally conceived, especially once a project has reached a stage of reasonable completion.