Google Summer of Code 2019

FOSSology: Continuation of Atarashi OSS

horizontal line

12August **20**19 - 18 August 2019 [ Week 12 ]

# Delivered By: Ayush Bhardwaj ([HastagAB](https://github.com/hastagAB))

Mentored By:

* Anupam Ghosh ([ag4ums](https://github.com/ag4ums))
* Gaurav Mishra ([GMishx](https://github.com/GMishx))
* Aman Jain ([amanjain97](https://github.com/amanjain97))

**[ Last Week ]**

# Goals for the Week

1. Implementing new algorithm
2. Testing and Fixing Bugs
3. Documenting and wrapping up project

# Tasks Accomplished

1. **Implementing new algorithm**

The machine learning model was successfully trained on the dataset provided by spdx. Loading of the model along with the text pre-processing was needed to implement Semantic Text Similarity algorithm for open source license scanning. The algorithm is implemented correctly as planned. Although the result as of now is not so accurate, as the model is trained on very smaller and poor dataset(~400 files with full license text). Also the text-preprocessing is done only using python in-built functions.

There is a huge scope of improvement in the algorithm. With the dataset of(10,000 files atleast) will make the model more accurate. The perfect text preprocessing can be achieved using SpaCy Library.

The speed of the algorithm is amazingly fast as compared to others. Also, If we scan a file, the correct license name is in the top 10 result list(if not on top)[according to sim score] . So it's a good sign. Improving the accuracy in the future will give us the most accurate and best results and will make Atarashi more powerful & faster than ever.

2. **Testing and Fixing Bugs**

The algorithm evaluator script was tested and few small bugs were fixed. The changes were done and the PR is approved and ready to be merged.

Link to the PR: <https://github.com/fossology/atarashi/pull/57>

For the new algorithm, before creating the PR, I manually tested it for different files and improved the code. I have done various testing for every small features in the codebase and after being fully satisfied with it, I created the PR. I have also optimized the codebase to be so compact and powerful as much as I can. Reduced the code size and improved the performance for the implementation.

Link to the PR: <https://github.com/fossology/atarashi/pull/58>

For the integration part of Atarashi in fossology, the installation script was needed to be created for different linux distros. The installation script includes Python, pip and Atarashi installation in the system using pip.

Link to the PR: <https://github.com/fossology/fossology/pull/1408>

3. **Documenting and wrapping up project**

The documentation inside the code for sphinx was done in atarashi. Documenting the whole GSoC period with all the contributions, works, discussions and future work was done. I’m happy to successfully achieve what I proposed with some extra features. For this I’m so thankful to all my mentors for helping and motivating me from time to time.   
I’m looking forward to continue contributing in fossology especially Atarashi even after the GSoC period.

# Conference Call

**17th August 2019(Saturday) Timings : 6:00 PM onwards**

**Attended By:** Ayush Bhardwaj, Sandip Bhuyan, Vivek Kumar, Michael Jaeger (Mentor), Anupam Ghosh (Mentor), Aman Jain (Mentor)

This was the last status update call. Everyone discussed their whole work done so far for GSoC and future plans. I discussed the complete work done so far and the few changes which will wrap up my GSoC project.