Google Summer of Code 2019

FOSSology: Continuation of Atarashi OSS

horizontal line

05August **20**19 - 11 August 2019 [ Week 11 ]

# 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))

# Goals for the Week

1. Coding the New Algorithm: Semantic Text Similarity

# Tasks Accomplished

1. **Coding the New Algorithm: Semantic Text Similarity**

Firstly trained the model with some inbuilt few texts and then with the common datasets floating around the web. These datasets included Google News dataset, Lee corpus, IMDB datasets and some countries dataset provided by tensorflow. But all of them were giving rubbish results. So needed to find some other dataset that will give some better results for license statements.

Found the spdx list of license statements containing all the licenses in the .txt format.  
<https://github.com/spdx/license-list-data>

Some of the statements are really long and need lots of text manipulation but I will be able to train the model on that. I’m hoping this will give better results than any other dataset available.

After the training, most of my work will be completed and only finding the cosine sim will be left that can be parsed as the result.

# Conference Call

**There was no conference call this week as I had to attend some other event on the day call usually schedules.**