CSCI 4961 Open Source Software 2019 Summer

Analysis of Open Source Projects

Robin Hong, hongz@rpi.edu

Brief Analysis

I have picked three projects for my open source projects analysis, based on my personal interests and their influence on me:

- **Submitty** from Rensselaer Center for Open Source Software.
- **gRPC** from Google Open Source.
- **TensorFlow** from Google Open Source.

Submitty is an open source programming assignment submission system hosted on GitHub, launched by the Department of Computer Science at Rensselaer Polytechnic Institute.

Evaluation Factor	Level	Evaluation Data
Licensing	2	BSD 3-Clause License is used (GitHub license page).
Language	1	My preferred language is C, C++, Python, and Java. The project uses C++, Python, and JavaScript extensively.
Level of Activity	2	It is really active in all quarters of last year.
Number of Contributors	2	The project has 92 contributors in total.
Product Size	2	It has much more than 10,000 lines of code.
Issue Tracker	2	Currently it has 283 open and 1,421 closed issues. Frequent debuggings are observed.
New Contributor	1	There is only a little bit of evidence of welcome or instructions for new contributors in the front page (how to contribute).
Community Norms	0	There is no evidence of documented and easy to locate statement of community norms that is welcoming and inclusive.
User Base	2	**Submitty** is heavily used in RPI and is definitely considered to have an active and engaged user base.
Total Score	14	

gRPC is a modern RPC framework that can run in any environment. It can efficiently connect services in and across data centers with pluggable support for load balancing, tracing, health checking and authentication. Currently it is used for communication in internal production, on Google Cloud Platform, and in public-facing APIs.

Evaluation Factor	Level	Evaluation Data
Licensing	2	Apache License 2.0 is used (GitHub license page).
Language	2	The project has mostly C code, which is my favorite.
Level of Activity	2	It is really active in the recent 12 months.
Number of Contributors	2	It has 465 contributors in total.
Product Size	2	It has much more than 100,000 lines of code.
Issue Tracker	2	It has 1,024 open and 6,220 closed issues. There is 6 issues reported in recent 3 days.
New Contributor	2	There is a welcoming and helpful note on how to contribute.
Community Norms	2	There is some sort of community norms (concepts and troubleshooting guide).
User Base	2	**gRPC** is used among many companies, including Netflix and Cisco.
Total Score	18	

TensorFlow is an end-to-end open source platform for machine learning, with a comprehensive and flexible ecosystem of tools, libraries, and community resources that lets researchers push the state-of-art in ML and gives developers the ability to easily build and deploy ML-powered applications.

Evaluation Factor	Level	Evaluation Data
Licensing	2	Apache-2.0 is used.
Language	2	It is mostly written in C++.
Level of Activity	2	It is really active in recent years.
Number of Contributors	2	It has 2,037 contributors in total and an average of 200 contributors per month.
Product Size	2	It has 2.5 million lines of code.
Issue Tracker	2	It has 2,287 open and 16,109 closed issues.
New Contributor	2	How to contribute is a conspicuous place and has welcoming notes and instructions for new contributors.
Community Norms	2	The community norms is also clear to see.
User Base	2	There is a clear evidence of heavy usage around the world.
Total Score	18	

In-Depth Analysis

Building from the knowledge I have on these three projects, I would like to select **TensorFlow** for my in-depth analysis. Last semester I was taking a CSCI course, Natural Language Processing, and the library of tensorflow is used extensively in the class, which has impressed me a lot.

TensorFlow's core open source library is designed to help us develop and train machine learning model. My chance of using TensorFlow is programming in Python. Running on my local machine take exceptionally long time to finish, which indicates that there is substantial workload of computing taking place inside TensorFlow.

Actually, TensorFlow has many derivative versions, like TensorFlow.js (JavaScript library for browser and Node.js), TensorFlow Lite (for deploying models on mobile and embedded devices), TensorFlow

Extended (end-to-end platform for preparing data, training, validating, and deploying models in large production environments).

Below is my extended analysis table for **TensorFlow**.

Evaluation Factor	Level	Evaluation Data
Licensing	2	**TensorFlow** uses Apache License 2.0 (GitHub license page), where commercial use, modification, distribution, and private use are allowed. Unlike copyleft license, it does not require a derivative work of the software or modifications to the original to be distributed using the same license. However, trademark use or liability is not given.
Language	2	For the statistics of languages, it has 51% of C++ code, 37% of Python code, 8% of HTML code, and 4% of others. Personally I prefer C++ and Python, and working with project like this type will be ideal: so I give 2 points here.
Level of Activity	2	It is really active in recent years. More specifically, it has 22,939 commits within 12 months from 901 contributors, and 1,887 commits within 30 days from 206 contributors. It is clear that this project is really prosperous and have been under active development for recent years. One thing worth to mention here is that **TensorFlow** has the very beginning commit in November, 2015.
Number of Contributors	2	It has 2,037 contributors in total and an average of 200 contributors per month. Moreover, the trend is that number of contributors is increasing steadily over months.
Product Size	2	Two years ago, the project only has 887,377 lines of code. Currently, the number stays above 2,297,782. This bulky volume of codes is definitely worth 2 points here.
Issue Tracker	2	It has 2,287 open and 16,109 closed issues. And in the GitHub page, there is detailed instructions on what to do when a bug is found. **TensorFlow** is quite considerate in issue tracking.
New Contributor	2	How to contribute is in a conspicuous place and has welcoming notes and instructions for new contributors. Generally speaking, it is a project friendly to new contributors.
Community Norms	2	The community norms is also very easy to see. It is divided into three categories: support, forum & user groups, and contribute. I can find what I want easily following the link. It also supports most media like blog, youtube, or twitter to get us informed.
User Base	2	There is a clear evidence of heavy usage around the world. Its fields of study include deep learning, neural network, GPU computing, machine learning, and numerical computing.
Total Score	18	This would be a perfect score.

Generally, the goal of TensorFlow is to provide handy and useful toolkit for computation in the field of machine learning and artificial intelligence. Lots of models of computing can be practiced by TensorFlow. And as a super huge open source project, it serves as a powerful and competent toolkit for lots of things.

Airbus is using TensorFlow to extract information from their satellite images and deliver crucial insights to their clients. Airbnb improves the guest experience by using TensorFlow to classify images and detect objects at scale. Tons of cases can be found with TensorFlow.

There is no way you can dwarf TensorFlow in next few years, where AI technology is supposed to develop quickly. It is appealing to me with its seemingly omnipotent power of computing, but also somewhat daunting to me because it is simply too big and too complicated for an undergraduate student to contribute something.