Proposal for CS 445/545 Final Project

Alex Brelsford[1], Savannah Norem[2], Divyani Rao[3], and Ty Vaughan[4]

I. INTRODUCTION

Where do you get the majority of your software from? You can now get most of the software you need to run your computer on the internet. One of the new trends in technology is contributing to open source software. Open source software is free, easily available ,and may be redistributed and modified.

There are multiple individuals and organizations such as universities, government research labs and companies that contribute to the open source world. These open source projects can be used by others to learn and improve. The reason to do this is to connect and create a community of developers who want to share their ideas and work.

In our final project, we want to explore the relationship between open source and where it comes from. There are multiple companies that are using the open source platform to distribute their software. In this project, we will be looking at certain projects that are popular in the open source world. We then want to answer two questions by looking at them: one) what are the companies that contributed to these projects, and two) which parts of the world (countries) contribute to these projects?

II. MOTIVATION

As a developer it is beneficial to know which areas are home to other people of similar interests. The world of open source software is vast and widely used by developers all over the globe, but certain areas create more open source projects than others. Developers are creating their own projects daily and it is difficult to see if developed countries are creating a greater portion of the open source world compared to undeveloped nations.

[1]A. Brelsford is a CS student in the EECS College of Engineering at UTK. alexbrelsford at gmail.com

[2]S. Norem is a CS student in the EECS College of Engineering at UTK. snorem1 at vols.utk.edu

[3]D. Rao is a CS graduate student in the EECS College of Engineering at UTK. drao at vols.utk.edu

[4]T. Vaughan is a CS graduate student in the EECS College of Engineering at UTK. wvaugha2 at vols.utk.edu

We want to find data on the locations of developers to determine the countries where the majority of open source projects are coming from. Adding to that, there are companies creating large open source projects that are being used by a vast number of people in their own open source projects. Some of these are fairly well known while others may not be.

We want to find these and see how many open source projects companies are creating. We also want to see what companies are behind the various open source software. To sum things up, we want to see how much companies and countries are influencing the world of open source software.

III. SOURCES OF DATA

We will be collecting our data from three sources:

- Libraries.io: Libraries.io monitors millions of open source projects that we will use to determine organizations that are creating open source software. We will be picking projects from libraries.io and looking at their top project languages and top project licenses. The licenses will give us an idea of who is working on the project - is it an academic organization or a company.
- Stackoverflow API: Stack Overflow is home to millions of programming related questions and users. We will use this API to gather data on user locations to determine how much each country is contributing.The data recieved from libraries.io can then be used to query Stackoverflow API to check which locations these projects come from.
- GitHub API: Since GitHub is one of the largest open source platform, we chose to collect data for projects in GitHub. The API provides us with information on whether an organization or an user has contributed to the projects.

IV. MILESTONES

In order to evaluate both the distribution of developers over developed and developing countries and to identify which companies are contributing to the growing trend of open source development, our team has laid out a time-line of milestones that we wish to

achieve. Do note that the given dates will change as needed to fit the class's syllabus. The following section discusses these milestones in detail.

A. Identify Resources and Gather Data - October 11

Our first objective is to identify all resources that are needed to perform our different analyses and to begin using these resources to access the specific data that we will need. It is important that we begin our data collection as soon as our sources are identified so that we know if the data sources are satisfactory or not. We plan to finish this step by October 11th.

B. Data Analysis and Interpretation - October 27

Once we have gathered our data, our team will begin interpreting the information and performing different analyses to solve our two problems. In order to streamline this process and perform a more in-depth analysis, we will split into two teams of two students, one pair to work on the data for the distribution of developers and the other to work on the data for which companies are behind open source projects.

C. Conclusions - November 3

After all of the necessary interpretations on our data have finished, we will begin drawing our conclusions. This includes preparation of our data and analyses for presentation.

D. Final Report - November 10

By this date, our team will have written and submitted all final materials, including any reports or presentations for the project.

V. Expected Outcome

While we are hoping to find data that surprises us, we expect to see trends towards more open source contributions from wealthy countries. In an article that was recently published by StackOverflow, it is shown how different countries have certain programming languages that are more dominant than others. In the article, StackOverflow obtained the results by interpreting questions asking about a specific language as the asking individuals using that language.

In a recent article [1], programmers in higher income countries are more likely to be using Python and R. This is likely because more research is performed in wealthier countries, and these two languages are dominant for such purposes. Since c and c++ are lower level languages that are more likely to be used in very specialized cases, programmers with more advanced degrees are shown to be more likely to use these languages. Lastly, programmers from underdeveloped countries are shown to be slightly more likely to use PHP and dramatically more likely to use a PHP framework called CodeIgniter.

Based on this data, we expect developers and companies in developed, wealthier countries to contribute to open source projects significantly more than those in underdeveloped, less wealthy countries. Large, well-known companies, such as Microsoft, IBM, AT&T and RedHat, are already major contributors to quite a few different open source projects. Similar companies and the countries that these companies are located in are expected to be among the key contributors.

References

[1] M. Byrne. Coders in wealthy and developing countries lean on different programming languages. Online, 2017.