CSE 564 Visualization

Prof. Klaus Mueller

Visualization of top Open-Source projects on Github

Grp 62

112503844 Mallikarjuna Rao Budida

112504241 Satish Reddy Muddana

Report Date: May 22, 2019

1.Aim of this Visualization

The primary goal of this visualization is to study the most trending Open-Source projects in the software industry. Visualizing such a data could help industry outsiders understand the shape and direction software is heading towards.

GitHub is a web-based hosting service for version control using Git. Github, being the first choice of open source contributors, was the perfect choice of point of interest for collecting appropriate data for this visualization.

Underlying Assumption: It has been safely assumed that the trendiness of a project is a measure of the number of forks it receives in the Github repository. The higher number of forks for a project implies more are the number of developers showing interest in working on the repository, hence makes it a trendy project.

Analyzing the data of top forked repositories on Github would closely resemble the trend in all Open-Source projects in the Software industry as well.

2.Collecting the data

APIs Used: Github API, Google Geolocation

Github provides APIs to developers to tinker with data on public repositories on the site.

Issues Faced in data collection**:** Github only gives the top 1000 results in any of their API responses. This issue was overcome while using Search API by executing the API for small intervals of the fork_count query over a large number of increment counts so as to cover a range of fork_count from 0 to 1M.

Also, Github only has text address of its repository owners, so the addresses had to be translated to Map Coordinates to make sense. This was done taking the help of the Google Geolocation API.

Execution**:** Once the basic data of each repository was pinned down, a separate crawl thread was spawned to get granular information like topics, owner information, owner address.

APIs were run in an Nginx Server running the custom-written PHP application. The PHP application stored the data in a local MySQL database. Data Cleaning, data mapping was done taking help of MySQL visualization tool called SequelPro.

3.Structure and Count of Data

We were able to successfully retrieve the data for top **8773** repositories by fork count, each having at least 500 fork count. For each repository, the attributes obtained were

owner_type
: Description of what the owner is (eg. Company, Individual, etc.)

created_at
: When the repo was created

updated_at
: When the repo was last updated

Size
: Size of the repo

language
: Coding Language the repo is written in, majorly

topics
: What topics does the repository fall into. Eg. NLP, ML, Visualization, etc.

forks_count
: Num of users who have forked the repository

stars_count
: Num of users who have “starred” the repository

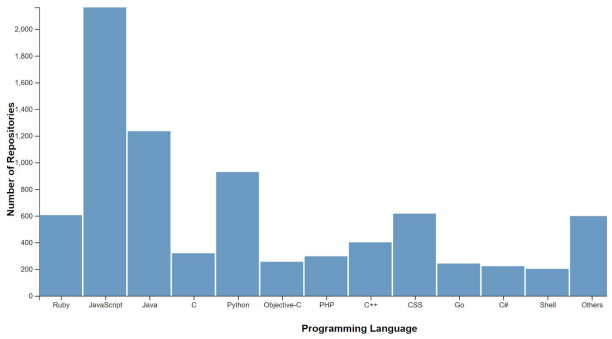
license
: Type of license the repo uses (MIT, Apache, etc.)

lat_of_owner, lng_of_owner
: Coordinates of the location the repo maintainer(s)/owner(s) are based out of

4. Types of Visualizations performed

Vis #1 Bar Graph

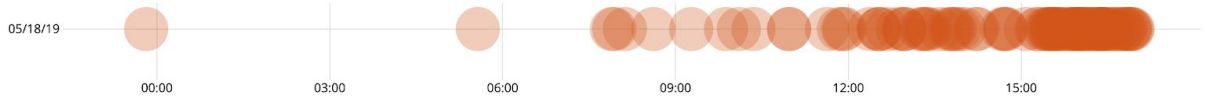
The languages vs repos implemented on that language were plotted as a bar graph



Javascript is the most preferred way of implementing open-source projects.

Vis#2 Time-series Plot

For a particular day(18 May 2019), the timestamps of commits were put on a time-series plot. A piece of very interesting information was inferred that Developers work the most from 3 pm to 12 pm.



All timestamps were normalized to GMT before analysis.

Vis#3 World Map with Bubble plot

The Lat/Long of the repository owners were plotted on a bubble World Map. It was observed that Silicon Valley (California) was where a major chunk of the projects was managed. Beijing, China was a close second.

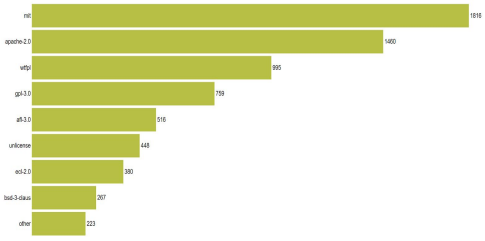
Vis#4 Word-Cloud

All topics associated with a repository(not to be confused with the language the repo is written in) were visualized as a word-cloud.

The shape of the word-cloud wrapper was decided to be the logo of Github as it has been shown to improve the effectiveness of communication (researched here).

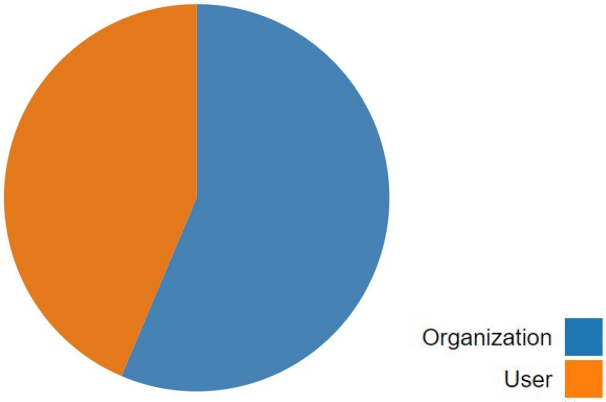
Vis#5 Horizontal Bar Graph

License distribution was plotted on a horizontal Bar Graph. MIT is the most followed open-source license.



Vis#6 Pie Chart

Owner type(Business Organisation or individual) was put to a pie-chart. These are the results



5.Inferences from the Visualization

Learning Javascript is the easiest way for programmers to contribute to open-source projects. Silicon Valley is where most of the contributors are located. Programming for Android, ios is the new trend. Organizations hold a major share of open source projects. It also looks like most of the work happens between 3 pm and midnight.

6.How this Visualization could help someone

As a young developer trying to get into open-source contribution, it might be overwhelming to learn where and what to put time and effort on. We tried to solve this problem by giving them an infographic of how open-source contribution is setup and which direction it is evolving.

7.Technologies used

DashBoard : HTML, CSS, Bootstrap, D3, JS, React
Data Collection & Cleaning : PHP, Python, MySQL

8.References

1. Github APIs(https://developer.github.com/v3/), Google GeoLocation API(https://developers.google.com/maps/documentation/geolocation/intro)
2. World Map Bubble visualization from https://www.react-simple-maps.io/bubbles-map
3. Word-Clouds generated by the tool on https://www.wordclouds.com/
4. Time-series plot from https://github.com/mlvl/timeseries
5. Tips on creating word-clouds https://medium.com/multiple-views-visualization-research-explained/improving-word-clouds-9d4a04b0722b