GitHub Analysis – Languages, Project Types, Users at Places

1-Term Project Synopsis

submitted to

by

Suraj Kumar Aavula
Reg. No: 181046037
Bigdata and Data Analytics

Naveen N Prabhu
Reg. No: 181046017
Bigdata and Data Analytics

Under the guidance of

Deepak Rao B

24th August 2018

INDEX

1. Abstract
2. Introduction
3. Methods
4. Results
5. Discussion
6. Conclusions

1.Abstract:

This project deals in extraction, processing, cleaning and storing the GitHub's Repositories data, generate the report to analyze the widely used programming languages, project types, and user locations. GitHub is a way to handle and maintain code for a developer in the sense that it keeps everything organized with the ability to undo any unwanted actions.

1.1. Keywords:

GitHub, Repositories, GitHub API, GH Archive, JSON

2. Introduction:

GitHub is a web-based 'version control system' which stores data in repositories, and allows other developers to access and modify the data stored in 'Git'. GitHub offers both public and private repositories for users to store their codes, public repositories are commonly used to host open source projects. As of June 2018, it has more than 28 million users and 57 million repositories (including 28 million public repositories). On June 4, 2018, Microsoft announced it had reached an agreement to acquire GitHub for US$7.5 billion. The purchase is expected to close by the end of the year.

API - Application Programming Interface, there are many open source libraries available for python to access the GitHub API. GitHub has open sourced its entire public repositories data for developers to access using API's and also the visualized data is available on google cloud here the data is updated per every hour, and the GitHub data available on Google cloud is explained on GH Archive. GitHub servers get thousands of hits per hour every day to make data available for everyone GitHub has restricted the users to access only 60 requests hour. The data retrieved is in the form of the JSON file which needs to convert to python so that it can be accessed like dictionaries.

3.Methods:

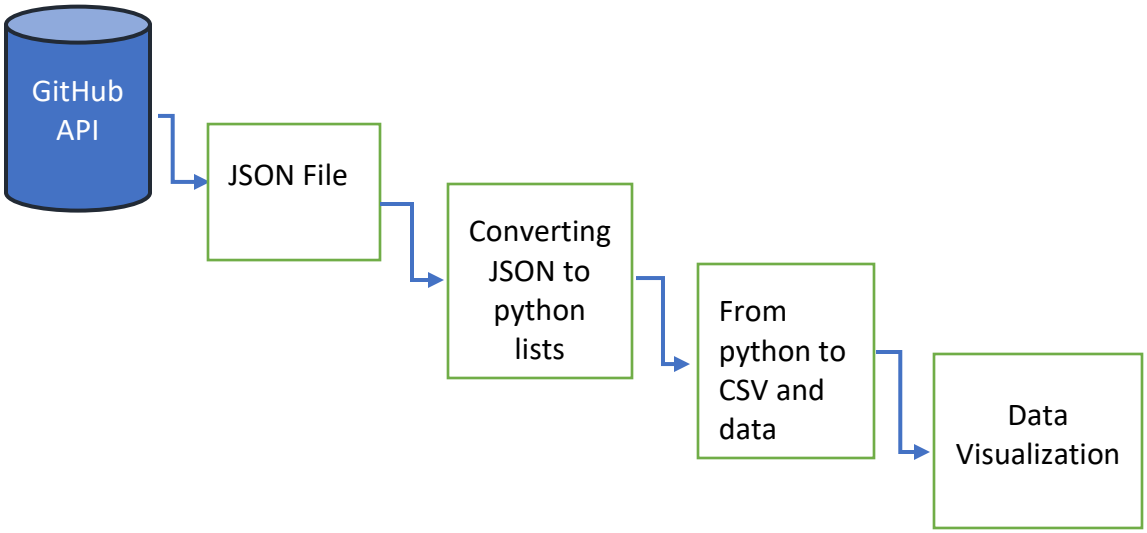
Open source developers all over the world are working on millions of projects i.e., is writing code and documenting, fixing and submitting bugs, and so forth. GH Archive is a project to record the public GitHub timeline, archive it and make it easily accessible for further analysis.

Each Archive contains JSON encoded events as reported by the GitHub API, raw data can be downloaded and processing can be done on it. Here in this project, we retrieve the data from the GH Archive using an API and we can process, clean and store the data to visualize.

The data has been extracted using GitHub API, the retrieved data is in a JSON format which has been converted to Python's dictionaries. After that, it is can be converted to CSV (Comma Separated Value) file. Here the only the required attributes have been taken while extracting the data, such that the cleaning of data is done by the time of retrieval. As the data is cleaned the analysis is done and the data has been visualized.

Block Diagram:

below block diagram showing the process of extraction and analysis of the data retrieved from GitHub



Below picture gives the understanding of how the user's data is organized on GitHub with help of API


```
{
  "login": "mojombo",
  "id": 1,
  "node_id": "MDQ6VXNlcjE=",
  "avatar_url": "https://avatars0.githubusercontent.com/u/1?v=4",
  "gravatar_id": "",
  "url": "https://api.github.com/users/mojombo",
  "html_url": "https://github.com/mojombo",
  "followers_url": "https://api.github.com/users/mojombo/followers",
  "following_url": "https://api.github.com/users/mojombo/following{/other_user}",
  "gists_url": "https://api.github.com/users/mojombo/gists{/gist_id}",
  "starred_url": "https://api.github.com/users/mojombo/starred{/owner}{/repo}",
  "subscriptions_url": "https://api.github.com/users/mojombo/subscriptions",
  "organizations_url": "https://api.github.com/users/mojombo/orgs",
  "repos_url": "https://api.github.com/users/mojombo/repos",
  "events_url": "https://api.github.com/users/mojombo/events{/privacy}",
  "received_events_url": "https://api.github.com/users/mojombo/received_events",
  "type": "User",
  "site_admin": false
},
```


Here the below pictures gives an idea of how the repositories of users are stored on GitHub with the help of API, the format in the picture is of JSON which will be retrieved.

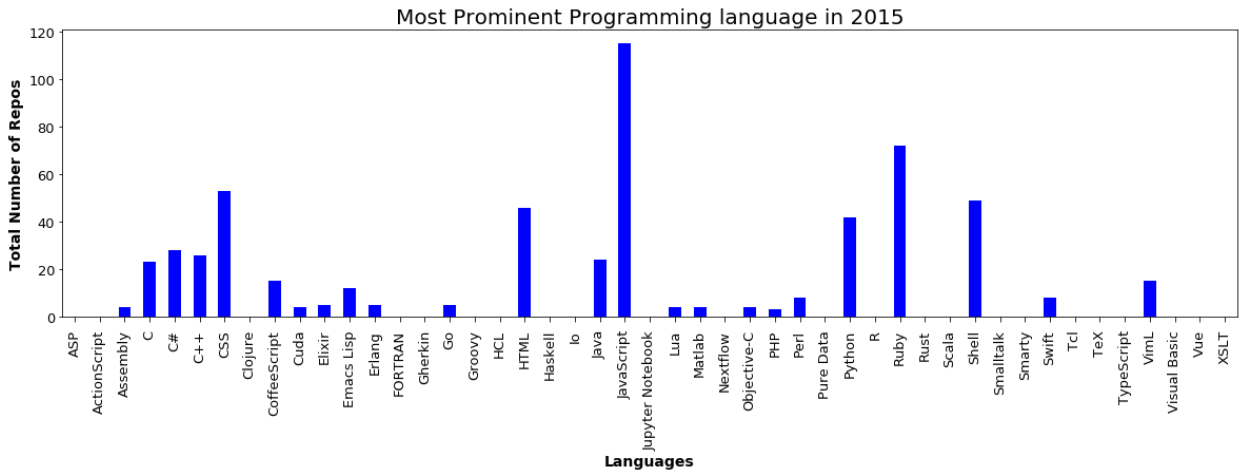

```
{
  "id": 26899533,
  "node_id": "MDEwOlJlcG9zaXRvcnkyNjg5OTUzMw==",
  "name": "30daysoflaptops.github.io",
  "full_name": "mojombo/30daysoflaptops.github.io",
  "private": false,
  "owner": {
    "login": "mojombo",
    "id": 1,
    "node_id": "MDQ6VXNlcjE=",
    "avatar_url": "https://avatars0.githubusercontent.com/u/1?v=4",
    "gravatar_id": "",
    "url": "https://api.github.com/users/mojombo",
    "html_url": "https://github.com/mojombo",
    "followers_url": "https://api.github.com/users/mojombo/followers",
    "following_url": "https://api.github.com/users/mojombo/following{/other_user}",
    "gists_url": "https://api.github.com/users/mojombo/gists{/gist_id}",
    "starred_url": "https://api.github.com/users/mojombo/starred{/owner}{/repo}",
    "subscriptions_url": "https://api.github.com/users/mojombo/subscriptions",
    "organizations_url": "https://api.github.com/users/mojombo/orgs",
    "repos_url": "https://api.github.com/users/mojombo/repos",
    "events_url": "https://api.github.com/users/mojombo/events{/privacy}",
    "received_events_url": "https://api.github.com/users/mojombo/received_events",
    "type": "User",
    "site_admin": false
  },
```

Below picture gives the details of the programming language used to develop an application stored in repositories.


```
{
  "Java": 12868,
  "HTML": 7582
}
```

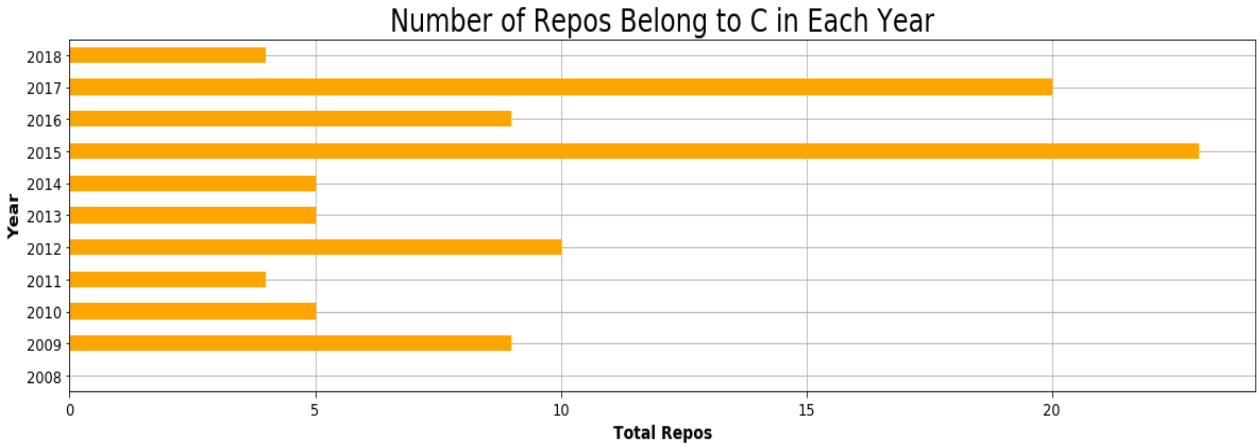

4.Results:

The analysis of the data retrieved has been done using the pandas library in python and the processed data has been visualized using the matplotlib library in python below figures gives the analysis report of the data which we retrieved from the GitHub.



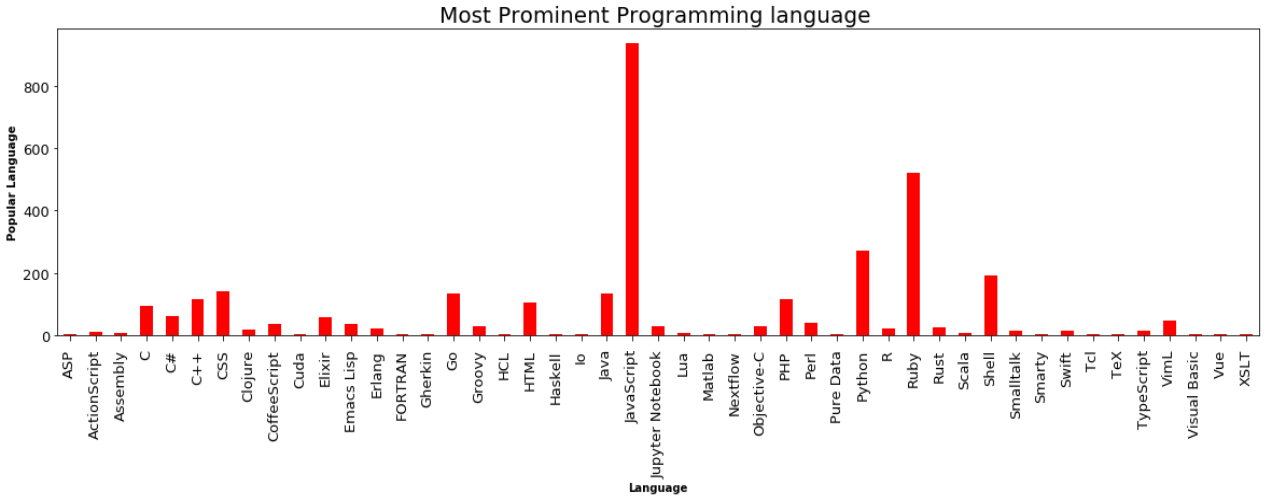
A most Prominent Programming language based on respective year

The above figure shows the most prominently used programming language by the developers in the market for the year 2015.



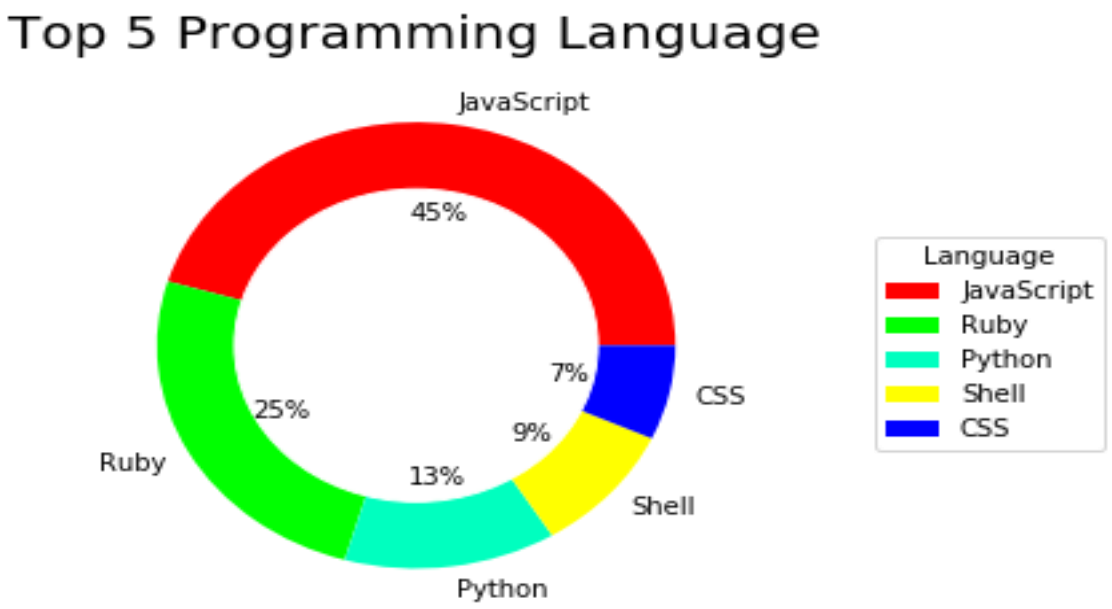
The popularity of the respective programming language across all the years

The above figure shows the popularity of C for all the years from 2008-2018 among the developers.

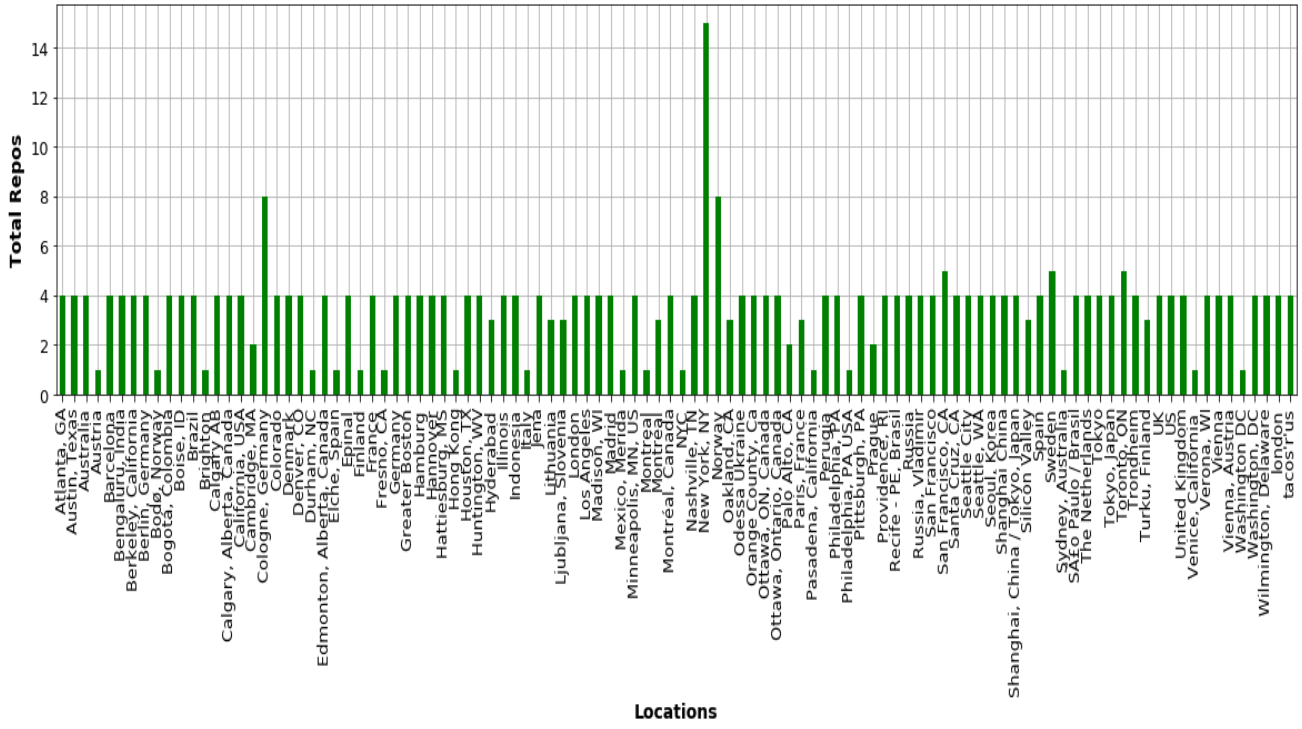


Most Prominent programming language among entire dataset

The above figure shows the most prominent programming language across all the years and all the repositories. As per our analysis JavaScript is the most prominent and popular programming language used by the developers across the globe.



The above figure shows the top five programming languages used by the developers. As per our analysis JavaScript takes a share of 45% among the popularity, followed by Ruby and Python.



Above figure show from which location wise accessing of GitHub by the developers across the globe.

5.Discussion:

As per the above results what we have obtained, we can infer that JavaScript is the most widely used language by the developers all across the globe. We have also analyzed the trend of the languages every year from 2008-2018 by using a bar chart, and also plotted the five most prominent programming languages in GitHub. JavaScript was the most prominent used language followed by Ruby and Python. The popularity of Python had been increased vastly in the last few years as the data processing has been trending and Python has a unique set of libraries to handle the large data sets. Python also has the popular machine learning libraries, It also has a special visualization library. These might be the reasons for the sudden growth in the popularity of the python.

As the traditional programming languages like C, C++, Java tends to dominate the developer's interest but lack to handle the rapidly growing customer demands, this makes the developers look out for other options which suitable to solve their problems.

Since the last 4-5 years, the popularity of python has been growing as never before as the demand for Big Data processing and the data analysis tools are growing rapidly. The people are also looking towards the most advanced technologies such as Machine learning, Statistical Modelling for analyzing the data, and python best fits to handle the huge data sets and also allows the developers to implement the suitable Machine Learning algorithms for predictions.

Ruby is one of the most popular languages among the startups. Its popularity is bolstered on Ruby on rails, a full stack web application framework that runs Ruby. Ruby has an incredibly simple beautiful syntax that allows a developer to do more with a less code. Ruby is a dynamically-typed language, which makes it very flexible and great for prototypes. These properties of Ruby makes it second most popularly used language among the developers in GitHub.

6.Conclusion

- GitHub is a way to handle and maintain code for a developer in the sense that it keeps everything organized with the ability to undo any unwanted actions.
- As per the objective of our project, the project dealt with the extraction, processing, cleaning and analyzing the GitHub repositories data.
- The analysis was done of the cleaned data which was extracted, the results were visualized using pandas libraries.
- JavaScript was found to be the most widely used programing language among the developers followed by Ruby and Python. (Results were also verified using other websites which showed the same analysis).