Assigning programming languages to geographic locations based on the activities of GitHub users

Gelera, Rodney
University of Victoria

McCulloch, Kaileen
University of Victoria

Warwick, Nick
University of Victoria

March 20, 2017

Contents

Abstract

Data has never before been generated at such high speeds. With the ever increasing volume of data becoming available every second, it has become more and more challenging to find meaningful information. Data mining and data visualization are two tools that can help solve this problem. In this paper we are going to use a classification hnique to simplify information and a visualizer as a quick and easy evaluation method. In order to explore these techniques we going to use GitHub user data to explore the use of programming languages according to geographic locations. We have scraped GitHub user profiles to get a location and a programming language for each user. From that we can classify each unique geographic location with a single programming language then our visualization tool allows us to explore the geographic distribution of language usage. There is a vast variety of development platforms available, selecting which tools to use can be a challenge. Whether you are a professional looking to stay in the game in this competitive climate or a person just starting their career in tech, looking at the language of choice in your area will help you make a more informed decision.

1 Introduction

GitHub is a popular online resource for software development.

2 Related Work?

3 Process

Tools used: GitHub, AWS CodeDeploy, Google Maps Geocoding API,

3.1 Data Retrieval

3.1.1 Downloading Data

The GitHub API returns JSON data with, among other things, user geographic location and programming language. Unfortunately, the API has a request limitation of 60 requests per hour. Each request return information for 30 users. In order to get 18000 user profiles it would take one IP 10 hours to download. This does not make for a fast or effective process.

3.1.2 Determining user location

GitHub allows users to next their location as a free-form text. This can be problematic for getting exact co-ordinates for a users location. Google Maps has a Geocoding API which converts text strings of geographic locations into latitudional and longitudional coordinates. Unfortunately this API has a usage limitation of 2,500 free requests per day.

3.2 Data Mining

The data retrieved in the *Data Retrieval* step needed to be converted into a geojson format into to be input into the *Data Visualizer*. Additionally, the user data needed to be aggregrated on a geographic location basis in order to classification each unique location with a single programming language.

3.3 Data Visualization

4 Conclusion

5 Future Work

6 Appendix A

6.1 Code