

Analysing commit messages

June 13, 2021

1 Introduction

With the rise of open source software, more and more big corporations incorporate free software in their stack. This also means that the amount of meaningful software that is available online on platforms like GitHub [1], is ever increasing. GitHub, as the name already suggest, offers the ability to host Git repositories. Git is a distributed version control system [2]. In this paper an effort has been made to analyse Git commits and gather information on their intended basic operation by looking at the message of the commit. Therefore repositories of the top 10 most wanted programming language according to [3] have been investigated. To get a more accurate representation of each language, 100000 commits have been chosen. To enhance diversity, only the last 10000 commits of each selected repository were used. The analysis has been done in Python.

2 Methodology

As already alluded to in section 1 the goal of the project is to analyse commit messages. To achieve this repositories of the top 10 most wanted languages. The reason for using this list as reference, is because it has a great assortment of languages that are in high demand and also relevant in today's development climate. The languages in question are:

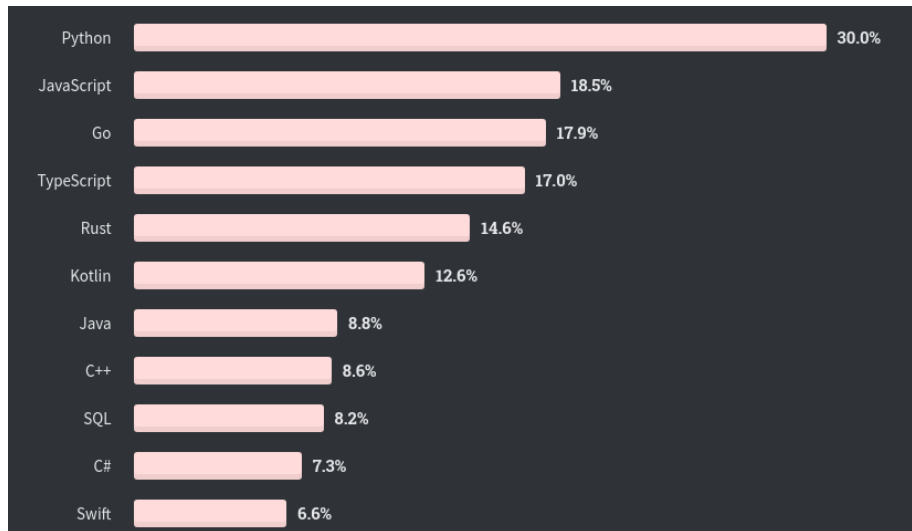


Figure 1: top 10 most wanted languages [3]

It has to be noted that not all languages depicted in Figure 1 have been chosen since SQL is not suited for analysis and therefore Swift has been chosen instead. Regardless of the percentages seen in Figure 1, 100000 commits of each language have been considered for analysis, where 10000 was the maximum per repository. The language of choice for the programming part of the analysis is Python. Python is well suited for this project as it has rich support for natural language processing related tasks. Additionally the language is great for quick prototyping of ideas and with the help of tools like Jupyter Notebook, document with graphs and text can be created in no time.