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 meaning-full 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. To full-fill on the premise 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, each repository is limited to 10000 commits. The analysis has been done in Python. Python is well suited for this project as it has rich support for natural language processing related tasks.

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 have been chosen, which are:

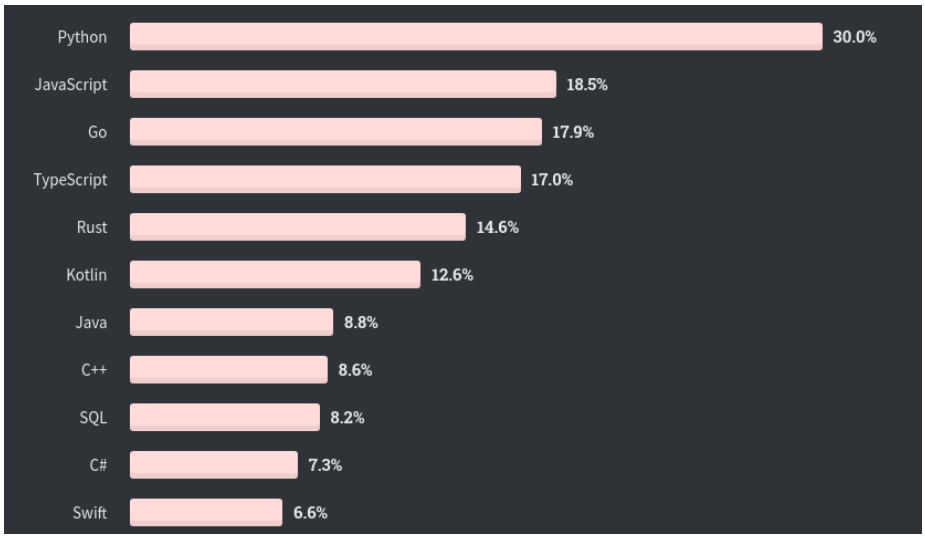

Figure 1: top 10 most wanted languages