How to Determine Similar Repositories in GitHub

Presenter: Jingci Wang

MS in Data Science

March 20, 2019

Motivation

Finding relevant projects is beneficial to developers in case of

- Reuse existing functions
- Explore ideas of possible features
- Analyze the requirements for their projects

Ideas

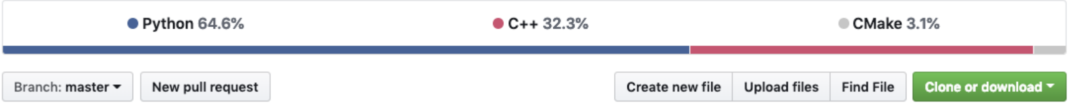
Based on GitHub users' history behaviors

- Similarity in users who starred same projects
- Similarity in users who watched same projects
- Similarity in users who pulled requests for same projects
- ...

Ideas

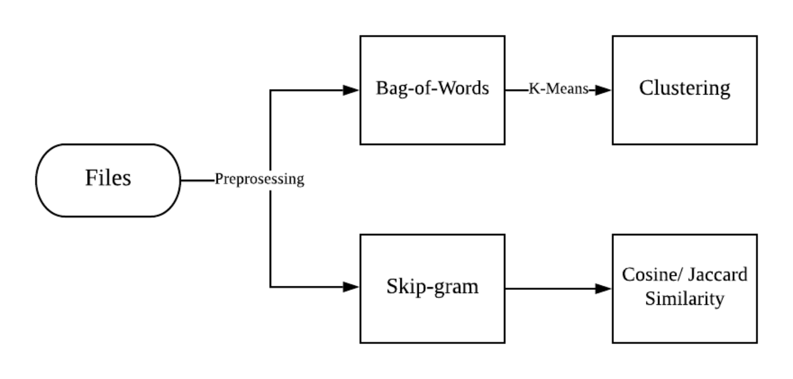
Based on Repositories contents

- Similarity in project description (Readme file)
- Similarity in Language Distribution


● Python 64.6% ● C++ 32.3% ● CMake 3.1%

- Similarity in Source Code

ReadMe File



- Not all repositories contain a ReadMe file

Source Code Identifiers

One of the richest features[1]

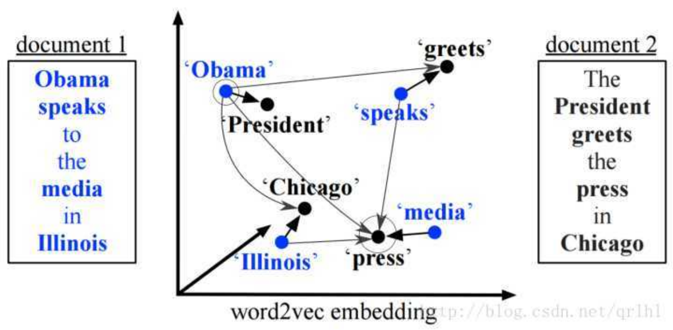
- Choose source code files: github/linguist
- Identifier extraction: pygments

$\Longrightarrow$ bag-of-words

[1] Topic modeling of public repositories at scale using names in source code

Pipeline

- **TF-IDF**(determine the weight of occuring identifier)
- **Topic model**
- **Embedding**
- **Nearest Neighbour** using Word Mover's Distance



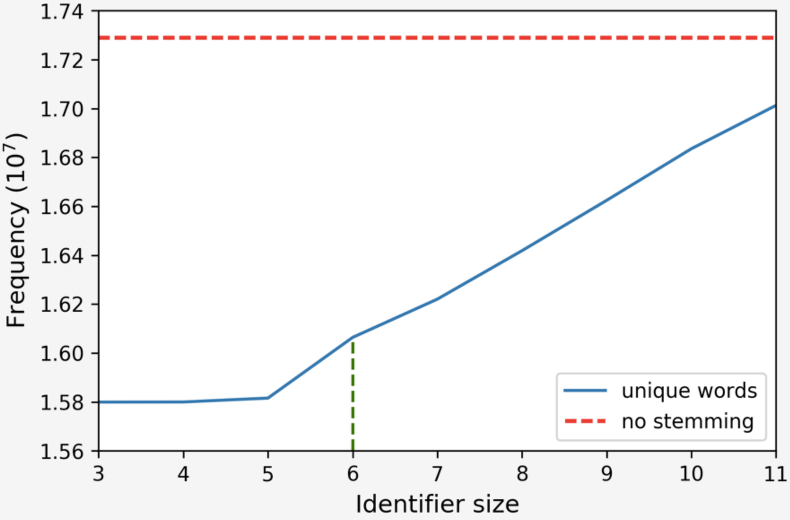
Extension

Hybrid Model

- Consider both description similarity and source code similarity

Potential Problems

- Giant Dataset
- Duplicated Repositories
- WMD has high complexity
- Different Natural Languages
- Curse of dimensionality


1.74
1.72
1.70
1.68
1.66
1.64
1.62
1.60
1.58
1.56
Frequency (10^7)
3
4
5
6
7
8
9
10
11
Identifier size
unique words
no stemming

Thank You