Bachelor / Master Thesis Description, status: assigned, language: [DE | EN]
Keywords: open source, license, github, data mining, python, quantitative data analysis

License Confusion on GitHub

Summary

Lopes at al. 2017 analyzed in their paper "DéjàVu: A Map of Code Duplicates on GitHub" source code of 4.5 million non-fork projects on GitHub, which contains 428 million files written in Java, C++ , Python, and JavaScript. They found out 70% of the code on GitHub consists of clones of previously created files.

In this thesis, the student will link clone-files to license information and estimate the amount or probability of license conflicts within projects on GitHub.

Knowledge in Python, statistics, and data analysis is a plus.

Details

- Literature review
  - Lopes at al. 2017: DéjàVu: A Map of Code Duplicates on GitHub
  - Basics of open-source licenses, their conflicts, license compliance
- Research approach
  - Analysis of the DejaVu data set and GitHub
  - Link clone-data with license information gathered from GitHub or from file headers
- Results
  - Analysis scripts written in Python, using Pandas and numpy
  - Probability/distribution of GitHub projects with a license conflict

Supervisor

Michael Dorner, michael.dorner@fau.de
Prof. Dr. Dirk Riehle, dirk.riehle@fau.de

Open Source Research Group
Computer Science Department
Friedrich-Alexander University

More information: http://osr.cs.fau.de/theses/resources/