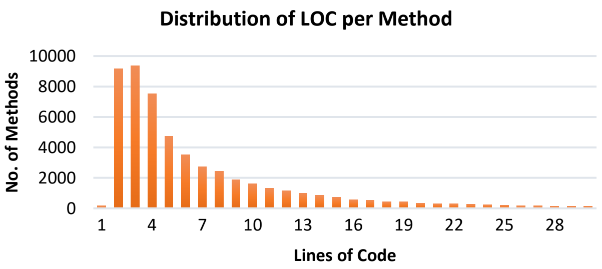
Source Code at Scale:
Analyzing Idiom and Pattern Usage across GitHub

Master Project Proposal, Summer Term 2017
Software Architecture Group, Prof. Dr. Robert Hirschfeld

Idioms, Patterns, and Code Metrics

Textbooks suggest a variety of means to improve code maintainability, including recommendations on *code metrics* (such as the length of methods), language-specific *idioms*, and larger structural and architectural blueprints – *design patterns* – for implementing recurring concepts. This project is concerned with a statistical analysis of code metrics, idioms, and patterns to track their real-world usage and identify factors which positively and negatively impact them.



Idiom Usage and Deviations	
`if x is not None:`	86%
`if not x is None:`	8%
`if x != None:`	5%
`if not x == None:`	1%

Example statistics

Mining GitHub

GITHUB DATA SET	
UPDATED	2017 Jan 01
NO. OF PROJECTS	39 703 095
NO. OF GIT COMMITS	502 481 630
NO. OF ISSUES	36 672 569
NO. OF PULL REQUESTS	16 826 727
NO. OF USERS	14 380 751
NO. OF FILE DIFFS	3 528 237 498 +

The initial data set, of which we provide an offline copy, is based on *GHTorrent*. This data set serves as a starting point to select representative projects, explore the full commit and collaboration graph, and avoid most round-trips to the GitHub API.

During the project, repository snapshots of representative projects at interesting timestamps as well as intermediate results, such as code metrics and extracted features, should be added to the data set.

Project Objectives

The two main objectives of the project are to provide statistical insights into code quality grounded in one of the largest data sets currently available and, by doing so, to explore strategies for dealing with large-scale source code processing and storage. As an intermediate goal, a representative and reproducibly selected sub-set of repositories for languages of interest should be built, along with fast parsers that create a code model sufficiently expressive to estimate simple code metrics and idiom densities. Basic knowledge in information retrieval, data mining, or statistics is useful but can also be acquired when needed.

Statistical Research Questions:

- In how far do code metrics, idioms, and patterns in real-world code match their textbook descriptions?
- Which factors (e.g. project size, programming language, number of developers, …) statistically explain variations in code metrics and idiom/pattern usage?

Technical Research Questions:

- Which disk/memory-based data structures for storing existing and computed knowledge about source code repositories support efficient retrieval of code metrics, idioms, and patterns?
- Which parsing and analysis strategies are feasible at terabyte scale?

Contact

Prof. Dr. Robert Hirschfeld, Toni Mattis, Patrick Rein
Software Architecture Group, Potsdam
http://www.hpi.uni-potsdam.de/swa, hirschfeld@hpi.uni-potsdam.de