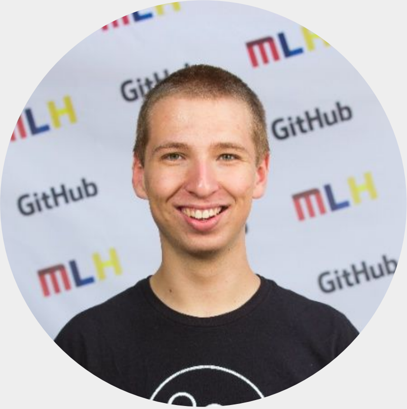
Open Source Analysis

Drew Casner



Carl Cortright

Shubha Swamy

Oliver Collins

Description

Our dataset contains information about open source projects. Using this dataset, we aim to track trends in programming languages, how popular repositories have changed over time, and contributions to them, and track repository life cycles.

Prior Work

- Companies that use this data
    - Tidelift: using open source data to make projects that utilize them more dependable
- Some work on analyzing Github repositories
    - Statistical analysis tools for git repositories
        - Gitinspector
    - Quality analysis on open source software

Datasets

- https://libraries.io/data
- 311 million data points
  - 2.7m unique open source packages
  - 31m repositories
  - 161m interdependencies

Proposed Work

- **Data cleaning/preprocessing/ integration**
    - **Removal of null values**
    - **Synchronize time zones**
    - **Match a unique user across multiple package managers**

List of Tools

- Data Analysis
  - Pandas
  - NumPy
  - SciPy
  - Jupyter Notebook
- Visualizations
  - Matplotlib
  - D3.js
  - Bokeh
  - graph-tool
  - Seaborn

Evaluation

- Evaluation will be based on a combination of the following:
    - A well written and interesting write up
    - Solid data that is mined with a strong analytic backing
    - Professional and interesting visualizations of our finding