Data flow

This file explains how the data and analysis were performed

1. Using the query “repo properties” the table repos2018 was generated. Table was exported to the file repos2018.
2. Using the Python file extract\_projects\_properties.py, the GitHub API was queried on the files repos2018. Results were saved to repos2018\_api.
3. Run build\_repo\_file.py to generate valid\_repos, repos\_full and statistics about file scope
4. Use queries from “File major extension” to generate the file repos2018\_major\_extensions
5. C repositories to exclude appear in repo\_exclusion (manual)
6. C\_repos is created using c\_projects.py
7. Tables c\_repose and valid\_repos (from the files) are uploaded to Bigquery to allow joining.
8. Run the ETL queries to build valid\_commits
9. Run Hotspots for file ccp analysis and hotspots
10. Run file\_hit\_rate\_dist.py and file\_ccp\_stability.py for the hotspots section
11. Use classify\_valid\_commits to label samples in valid\_commits and created valid\_commits\_classified
12. Create\_commit\_size - number of files (of different types) in a commit
13. Create\_refactor\_stats - computes before/after statistics of refactors and inserts them into refactor\_stats