source{d}

Public Git Archive: a Big Code dataset for all

pga.sourced.tech

content

BIG CODE DATASET

- ‡ **182,014** GitHub top-starred repositories
- ‡ **248,043** Siva files
- ‡ **3** TB of code
- ‡ **54,5** million of files
- ‡ **15,941** million of Lines of code
- ‡ **455** distincts languages
- ‡ **289** distinct licenses

index file

Column name	Description
url	URL of the GitHub repository.
siva_filenames	Siva files which contain parts of that repository.
file_count	Number of files in default HEAD reference.
langs	Languages encountered in default HEAD.
langs_{byte,lines,files}	Byte, line, file counts per each language, in the same order as *langs*.
commits_count	Number of unique commits in the Siva files which refer to that repo.
branches_count	Number of references, tags excluded.
fork_count	Number of remotes in the referring Siva files.
{empty,code,comment}_lines_count	Number of empty, code, commented lines in the default HEAD
license	License names and corresponding confidences.

description

LANGUAGES

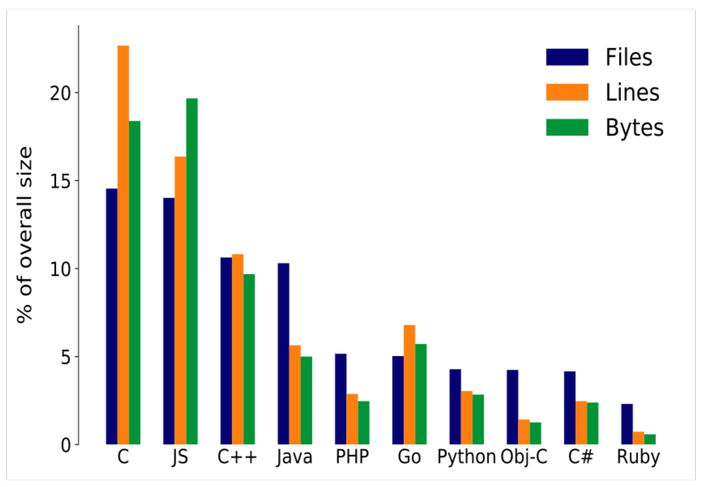

Figure 1: Statistics of 10 most popular languages in PGA

CODE VS. COMMENT

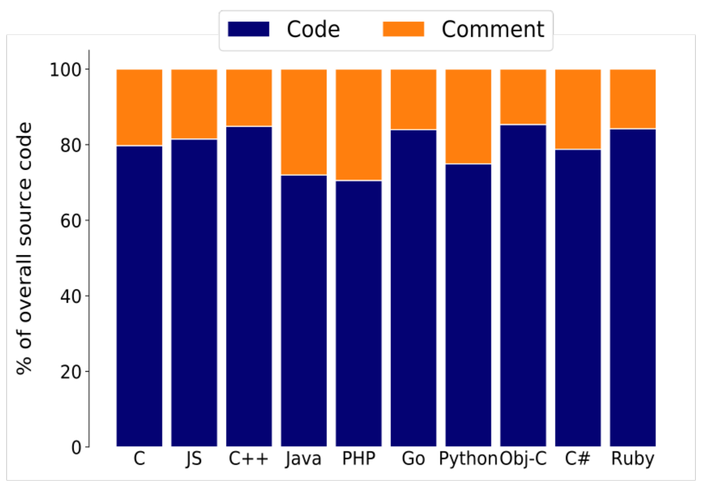

Figure 2: Proportion of lines of code and comments for the 10 most popular languages in PGA

usage

EXTRACT IDENTIFIERS FROM UAST

```
from sourced.engine import Engine
engine = Engine(spark, "/path/to/siva", "siva")

engine.repositories.references.head_ref \
.commits.tree_entries.blobs \
.classify_languages() \
.filter('lang = "Python"') \
.extract_uasts() \
.query_uast('//*[@roleIdentifier]') \
.extract_tokens("result", "tokens") \
.select("blob_id", "path", "tokens")
```

TOP DEVELOPMENT REPOS

	GitHub repos	Branches
1.	JetBrains/intellij-plugins	59,321
2.	openstack/cinder	48,515
3.	servo/servo	43,022
4.	google/angle	38,327

	GitHub repos	Commits
1.	OpenChannelSSD/linux	735,972
2.	google/ktsan	535,479
3.	mirrors/chromium	260,366
4.	Azure/azure-content	243,185

description

ILICENSES

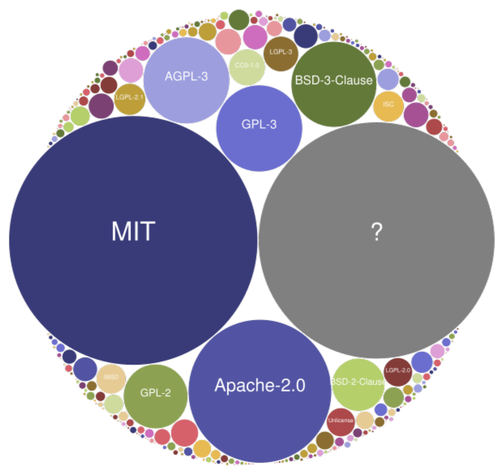

Figure 3: Distribution of licenses in PGA

BRANCHES AND COMMITS

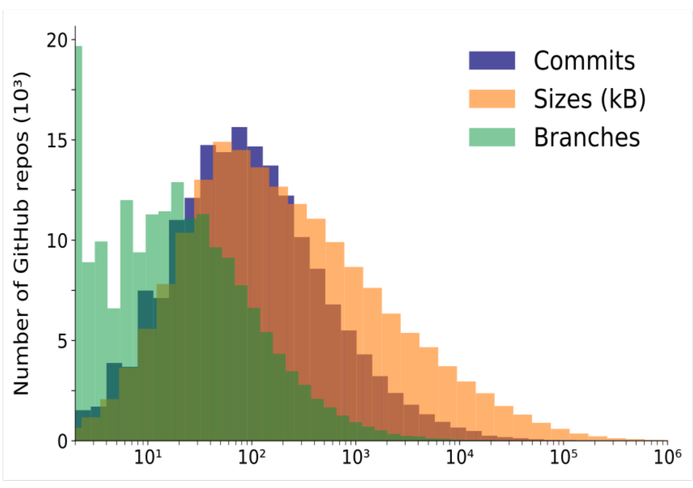

Figure 4: Distribution of the sizes and numbers of commits/branchesper repository in PGA

multitool

LIST AND DOWNLOAD THE REPOS IN PGA

- ‡ `discover` - extract the list of repositories from GHTorrent MySQL dump.
- ‡ `select` - compile the list of repostitories for cloning.
- ‡ `get-index` - download the index file of the latest dataset generated.
- ‡ `get-dataset` - download the Siva files from the specified list.

applications

ML ON CODE USING PGA

- ‡ Source code analysis and language modeling
- ‡ Identifier embeddings
- ‡ Topic modeling
- ‡ Code summerization
- ‡ Clone detection
- ‡ Program repair and bug detection
- ‡ Code suggestion and completion