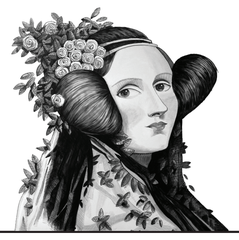
source{d}



Public Git Archive: a Big Code dataset for all

pga.sourced.tech

content.

BIG CODE DATASET

‡ **182,014** GitHub top-starred repos

‡ **248,043** Siva files

‡ **3** TB of Git repositories

‡ **54,5** million files in HEAD

‡ **15,9** billion lines of code in HEAD

‡ **455** distinct programming languages

‡ **289** distinct licenses

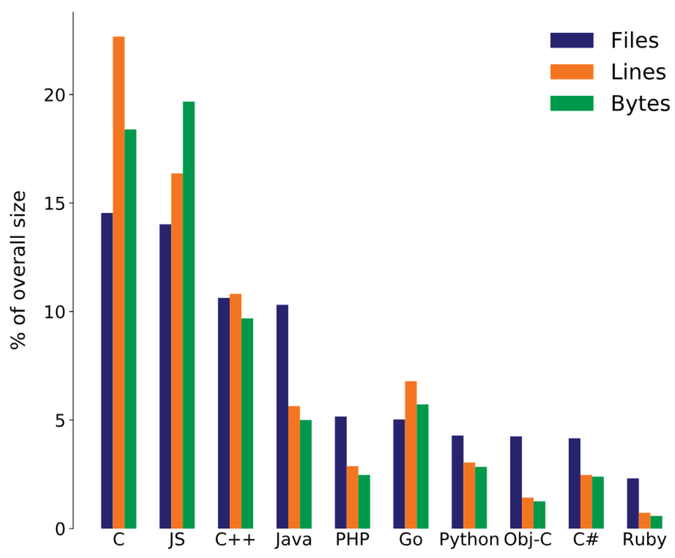
index file.

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



Statistics of the 10 most popular languages in PGA

usage.

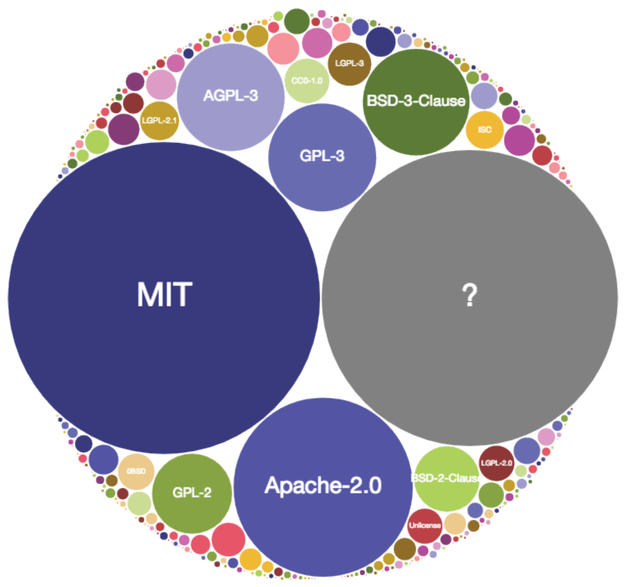
EXTRACT IDENTIFIERS

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

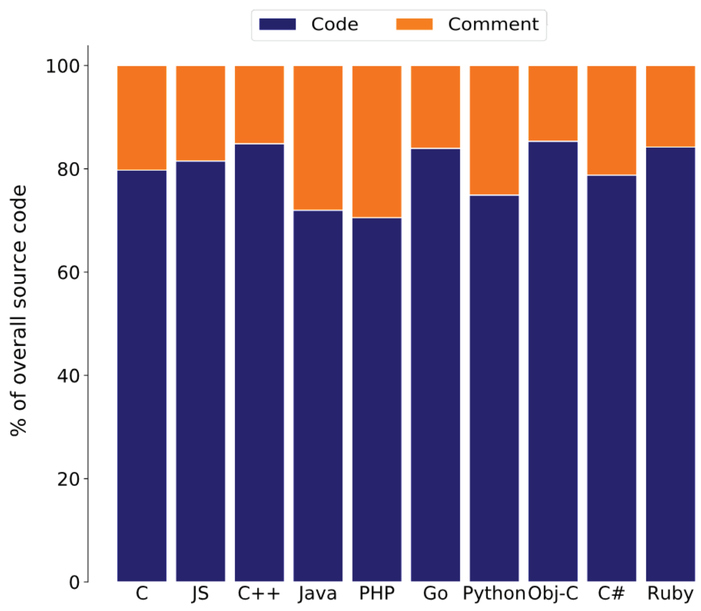
description.

LICENSES



Distribution of licenses in PGA

CODE VS. COMMENT



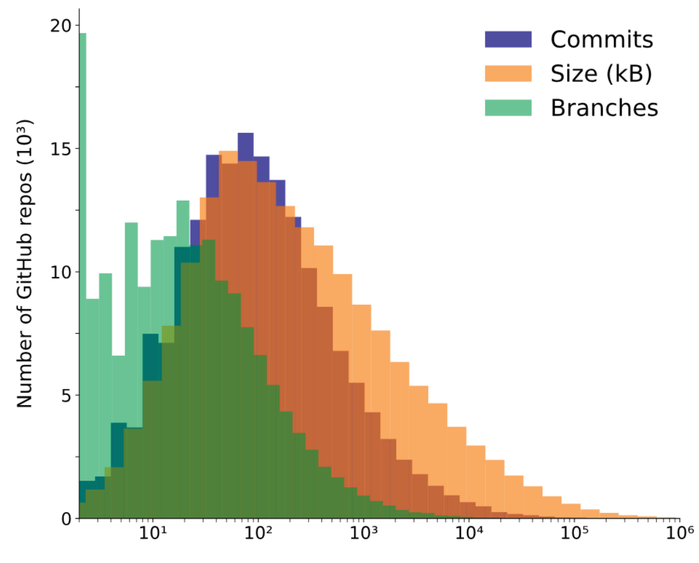
Proportion of lines with code and comments for the 10 most popular languages in PGA

BRANCHES AND COMMITS

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

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

BRANCHES AND COMMITS



Distribution of the sizes and numbers of commits/branches per repository in PGA

pga tool.

LIST THE REPOSITORIES IN THE INDEX

‡ `pga list`

`--format csv` to print CSV rows with all the details - supports also JSON

`--lang java,go` to list repos with at least some code in those two languages

`--url regexp` to list repos for which the url matches the given regexp

DOWNLOAD THE SIVA FILES

‡ `pga get`

`--output path` path where to store the siva files - supports HDFS.

`--jobs n` to set the maximum number of concurrent downloads - default 10.

applications.

#MLonCode USING PGA

‡ Source code analysis

‡ Identifier embeddings

‡ Topic modeling

‡ Code summarization

‡ Fuzzy duplicates detection

‡ Automatic program repair

‡ Code suggestion and completion