



### 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

siva\_filenames

file\_count

langs

url

langs\_{byte,lines,files}

commits\_count

branches\_count

fork\_count

{empty,code,comment}\_lines\_count

license

#### DESCRIPTION

URL of the GitHub repository.

Siva files which contain parts of that repository.

Number of files in default HEAD reference.

Languages encountered in default HEAD.

Byte, line, file counts per each language, in the same order as langs.

Number of unique commits in the Siva files which refer to that repo.

Number of references, tags excluded.

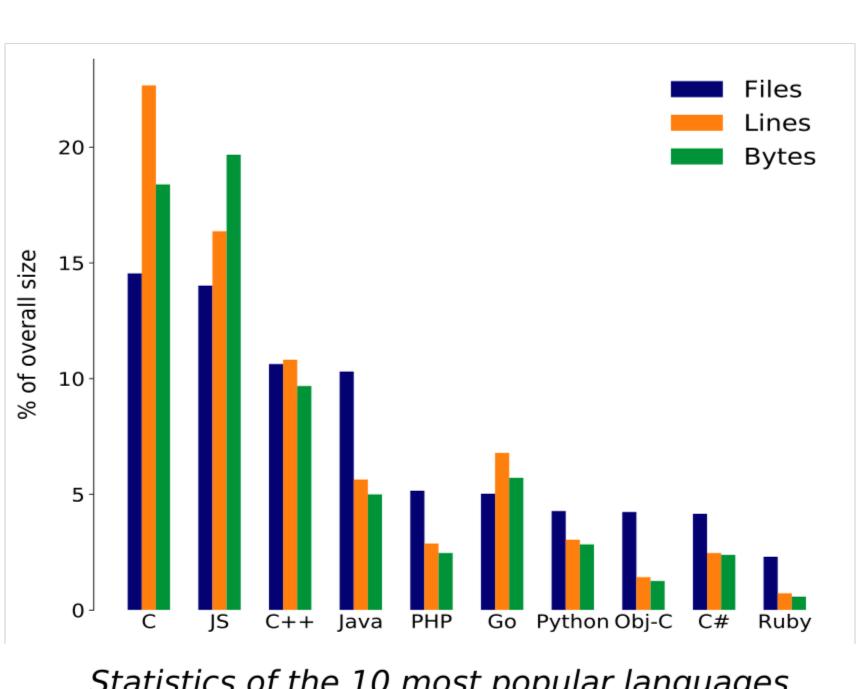
Number of remotes in the referring Siva files.

Number of empty, code, commented lines in the default HEAD

License names and corresponding confidences.

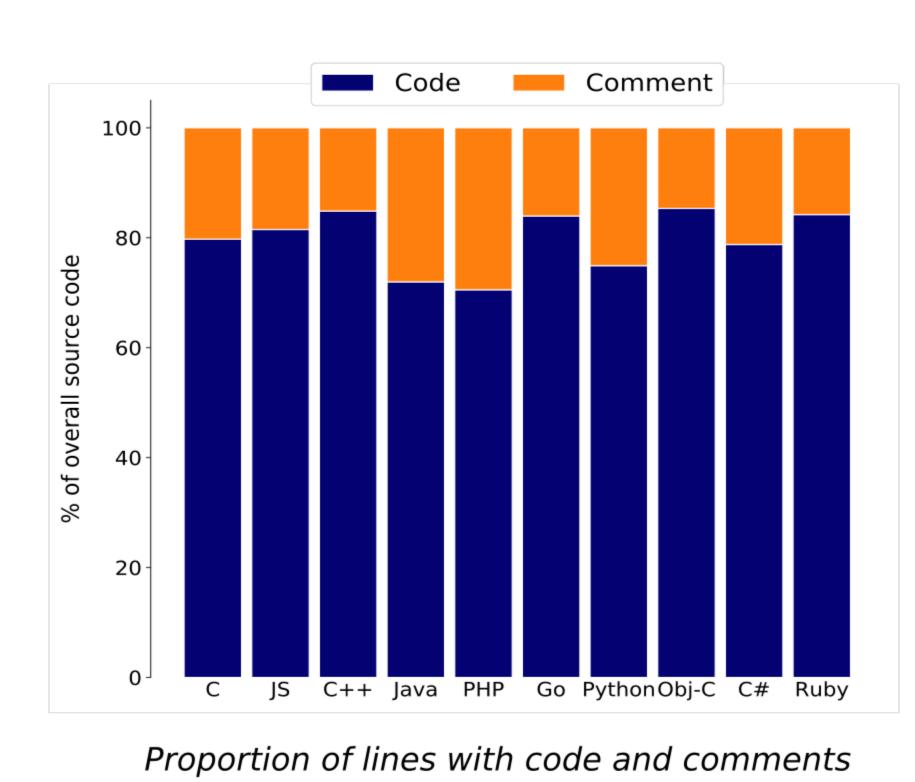
# description.

#### LANGUAGES



Statistics of the 10 most popular languages in PGA

#### CODE VS. COMMENT



for the 10 most popular languages in PGA

### usage.

#### EXTRACT IDENTIFIERS

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')

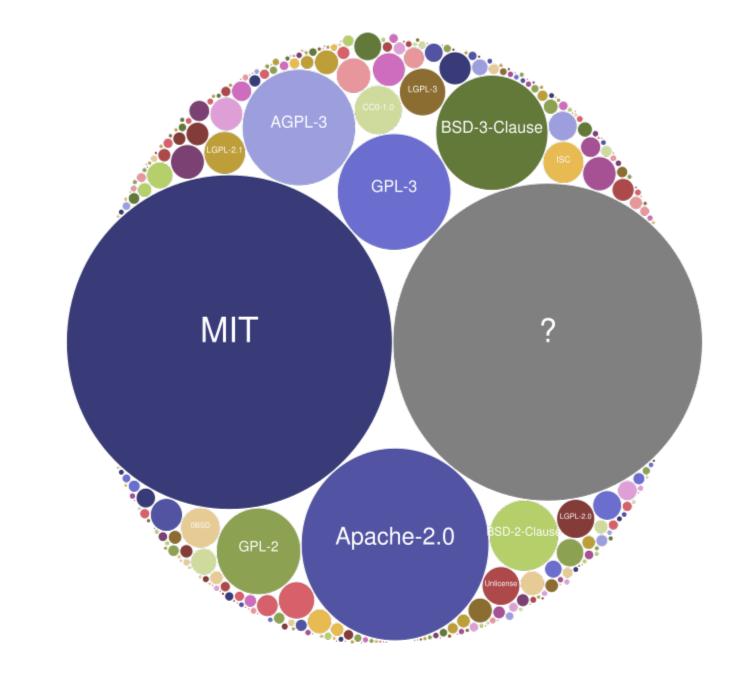
#### BRANCHES AND COMMITS

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

Re	epository	Commits
1. 0	penChannelSSD/linux	735,972
2. gd	ogle/ktsan	535,479
3. m	irrors/chromium	260,366
4. A	zure/azure-content	243,185

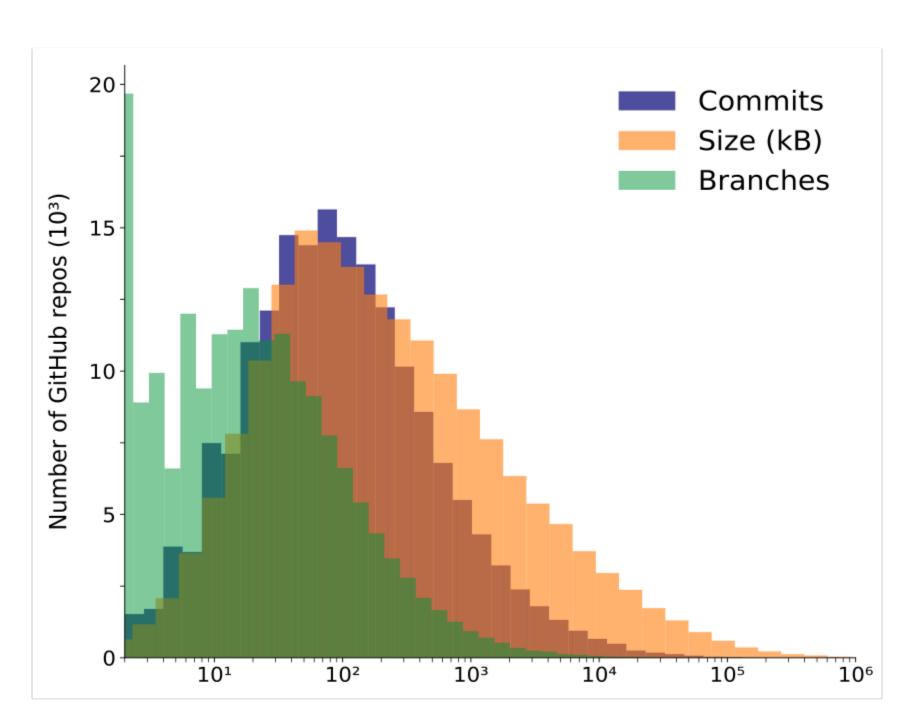
## description.

### LICENSES



Distribution of licenses in PGA

#### BRANCHES AND COMMITS



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

## pga tool.

#### LIST THE REPOSITORIES IN THE INDEX

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