More than half of GitHub is duplicate code, researchers find • The Register

By Richard Chirgwin 21 Nov 2017 at 03:57

Boffins beware: random samples are therefore useless for research

Given that code sharing is a big part of the GitHub mission, it should come at no surprise that the platform stores a lot of duplicated code: 70 per cent, a study has found.

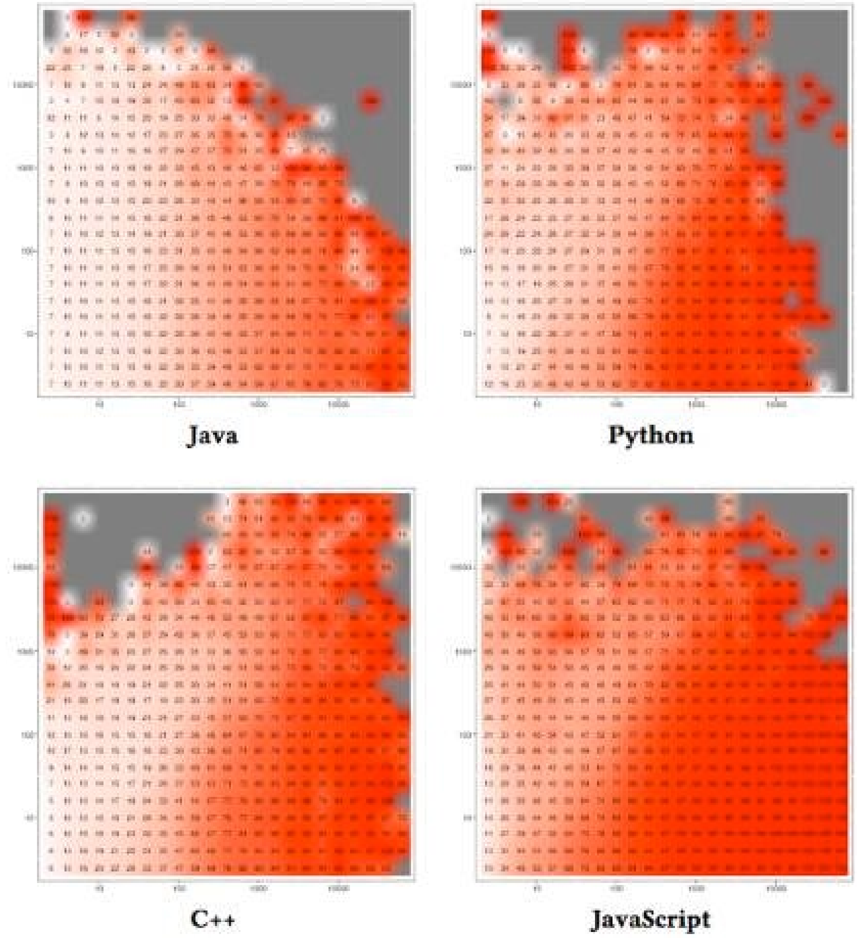
An international team of eight researchers didn't set out to measure GitHub duplication. Their original aim was to try and define the “granularity” of copying – that is, how much files changed between different clones – but along the way, they turned up a “staggering rate of file-level duplication” that made them change direction.

Presented at this year's OOPSLA (part of the late-October Association of Computing Machinery) SPLASH conference in Vancouver, the University of California at Irvine-led research found that out of 428 million files on GitHub, only 85 million are unique.

Before readers say “so what?”, the reason for this study was to improve other researchers' work. Anybody studying software using GitHub probably seeks random samples, and the authors of this study argued duplication needs to be taken into account.

As open source watcher Adrian Colyer blogged, “simple random selection is likely to lead to samples including high duplication, which may bias the results of research”, so the paper's resulting public index of code duplication, which they've dubbed “DéjàVu”, helps “understand the similarity relations in samples of projects, or to curate sample to reduce duplicates”.

For example, the study said, if a researcher is studying how many C and C++ programs use assertions, duplication clearly skews their output; similarly, a software quality study needs to take duplication into account.



Java

Python

C++

JavaScript

X = files, Y = commits, Color = dupes. Source: DéjàVu: A Map of Code Duplicates on GitHub, Lopes et al at ACM

DéjàVu maps file clones in Java, C++, JavaScript and Python.

The researchers assessed code duplication using a variety of hash techniques. Identical code was easy, since they produced identical hashes, but it was also necessary to take into account software with small changes (spaces or tabs), or even larger changes.

To draw these other duplicates into their sample, the researchers applied a “token hash” that captured minor changes in spaces, comments, and ordering; and a package called

```
SourcererCC
```

to capture clones with edits too large for the token hash.

JavaScript was the most cloned environment of all: a mere six per cent of files spawned the other 94 per cent of JavaScript files on GitHub. Of the C++ ecosystem, 73 per cent of files were duplicates; 71 per cent of Python programs were dupes.

Java developers are the most individualistic of the four environments researched, “but even for Java, 40% of the files are duplicates”.

The other thing that probably won't surprise readers is that duplication is primarily dependency-driven. JavaScript provided a good example: people creating a project would commit NPM libraries into their new repositories as if they were part of the application code.

As Colyer wryly noted: “If ever you have felt like you are downloading the universe when running `npm install`, here’s the data to prove it: including nested dependencies (nesting up to 47 levels deep was discovered, with median five) the number of unique included projects has median 63, and maximum 1261.”

Similarly, nearly all JavaScript programmers suck JQuery into their projects.

There's also programmers' habits as Git users: “there is a lot more duplication of code that happens in GitHub that does not go through the fork mechanism, and instead, goes in via copy and paste of files and even entire libraries”, the study noted. ®