82% of code on GitHub is copying existing files

That's the result of a recent study by researchers from the University of California, Irvine, Microsoft Research, Czech Technical and Northeastern.

They studied 4.5 million original projects on GitHub, containing all 482 million files.Of which only 85 million, equivalent to 17.63% is the original.

Most JavaScript projects contain copy files

The study only looks at projects written in C ++, Java, JavaScript and Python.In which JavaScript has 94% of the file is the same clone as the original (based on hash file).C ++ ranked second with 73%, followed by Python and Java, respectively 71% and 40%.The researchers also looked at the content of the file (based on token hash), but the results were similar.

Table 3. File-Level Duplication.

	Java	C++	Python	JavaScript
Total files	72,880,615	61,647,575	31,602,780	261,676,091
File hashes	43,713,084 (60%)	16,384,801 (27%)	9,157,622 (29%)	15,611,029 (6%)
Token hashes	40,786,858 (56%)	14,425,319 (23%)	8,620,326 (27%)	13,587,850 (5%)
SCC dup files	18,701,593 (26%)	6,200,301 (10%)	2,732,747 (9%)	5,245,470 (2%)
SCC unique files	22,085,265 (30%)	8,225,018 (13%)	5,887,579 (19%)	8,342,380 (3%)

The result of the number of files copied on GitHub

The reason is NPM

NPM is a library management tool (package or module) for both client and server in JavaScript projects.NPM is currently the world's largest package management tool with over 350,000 libraries, more than double the second tool - Apache Maven.

NPM contains many useful libraries so many developers use it.Therefore, they import more JavaScript project libraries than other languages, and the number of reusable code is also more.

See also: Microsoft and GitHub cooperated to bring Git virtual file system to macOS and Linux

Research on code is important for other studies

'Git, the source control system on GitHub is built to encourage' taking '(copying) the project.But many of the code is copied without a grab and copy file and even the entire library. '

This research result is important because 'firstly, maybe GitHub should recapture its real data scale and secondly, more and more research using a huge number of open source projects is available on GitHub. .Copying such code will affect the results of these studies.The raw data for this study can be downloaded here. *http://mondego.ics.uci.edu/projects/dejavu/* This is the whole study. *http://janvitek.org/pubs/oopsla17b.pdf* and *https://dl.acm.org/citation.cfm?doid=3152284.3133908*