DéjàVu: a map of code duplicates on GitHub

Typ	Tidskriftsartikel
Författare	Cristina V. Lopes
Författare	Petr Maj
Författare	Pedro Martins
Författare	Vaibhav Saini
Författare	Di Yang
Författare	Jakub Zitny
Författare	Hitesh Sajnani
Författare	Jan Vitek
Webbadress	http://dl.acm.org/citation.cfm?doid=3152284.3133908
Band/Årgång	1
Nummer	OOPSLA
Sidor	1-28
Publikation	Proceedings of the ACM on Programming Languages
ISSN	24751421
Datum	2017-10-12
DOI	10.1145/3133908
Hämtad den	2019-09-02 16:26:39
Bibliotekskatalog	Crossref
Språk	en
Kort titel	DéjàVu
Tillagd den	2019-09-02 16:26:39
Ändrad den	2019-10-07 12:37:46

Etiketter:

Cloning, GitHib mining

Bifogade dokument

- Lopes m. fl. - 2017 - DéjàVu a map of code duplicates on GitHub.pdf

  [8 timmar]

  Undersöker mängden (i olika utsträckning) klonade kodfiler för alla C++-, Java-, Python- och JavaScript-projekt på GitHub: De går igenom alla projekt som inte är forkade från ett annat projekt och extraherar följande information för varje kodfil:

  - MD5-hash av hela filen
  - MD5-hash av alla tokens, inkl. information om frekvens
  - storlek i Bytes
  - #rader
  - #kodrader (med/utan kommentarer)
  - #token (totalt + unika)

  De jämför fil- och tokenhashar, och använder SourcererCC för att hitta filer som är minst 80% lika (men där hasharna diffar).

  De analyserar även metadata från GHTorrent.

  De kommer fram till att:

  - Java- och C++-projekt innehåller fler filer än Python- och JavaScript-dito.
  - C++-filer är större än filer från de andra språken, och JavaScript-dito är mindre.
  - De flesta projekt är små och varken särskilt aktiva eller populära.
  - De flesta filerna är kloner (har samma hashar) och endast följande andel är unika filer:
    - Java: 60%
    - C++: 27%
    - Python: 29%
    - JavaScript: 6%
  - Den mest duplicerade filen är en tom fil (0B). Den näst mest duplicerade innehåller en (1) tom rad.
  - De mest duplicerade filerna är generellt små. Om man exkluderar de minsta filerna får man dock bara marginellt lägre klon-frekvens.
  - 5% av Java- C++- och Python-projekten är exakta kopior av något annat projekt (modulo whitespaces, kommentarer och terminal-symboler). Motsvarande siffra för JavaScript är 11%.
  - Betydligt fler projekt än så överlappar till en stor del andra projekt.

  För 4 olika filstorlekar analyserar de de 20 mest klonade filerna inom varje språk, och noterar att:

- De mest duplicerade filerna kommer från ett fåtal välkända bibliotek och ramverk.
- Många repron som är beroende av något annat projekt P har kopierat in hela P:s källkod och komittat tillsammans med sin egen kod. Ofta innehåller P på samma sätt källkoden från projekt som det är beroende av.

Här är JavaScriptprojekten särskilt uppseendeväckande då många av projekten kallade P ovan finns tillgängliga via pakethanteraren NPM. P i sin tur inkluderar andra NPM-paket på samma sätt.

De analyserar 20 slumpvisa klonpar för att se om de är avsiktliga kloner eller inte. Resultaten ser olika ut för olika språk och är för långa för att redogöra för här. Se stycke 6.2.3.

Påpekar att den stora mängden kodkloner kan ge "biased" resultat, men forskare antar generellt att de olika observationerna (repositorierna) är oberoende av varandra!!!