# Code Cloning Habits Of The Jupyter Notebook Community

| | |
|---|---|
| **Typ** | Tidskriftsartikel |
| **Författare** | Ulf Sigvardsson |
| **Sidor** | 48 |
| **Bibliotekskatalog** | Zotero |
| **Språk** | en |
| **Tillagd den** | 2019-09-02 16:37:42 |
| **Ändrad den** | 2019-10-07 12:43:39 |

## Etiketter:

Cloning, GitHib mining, Jupyter

## Bifogade dokument

- Sigvardsson - Code Cloning Habits Of The Jupyter Notebook Commun.pdf

  Samlar ihop Jupyter Notebooks från github och extraherar alla kodsnippets, totalt 32 288 102 snippets från 2 603 321 notebooks. Han klipper bort alla kommentarer, blanktecken och radbrytningar. Därefter MD5-hashar han alla snippets och mappar snippets med identiska hashar (=typ 1-kloner) till ett och samma "kluster".

  Han räknar ut dupliceringsratio för varje notebook som:
  (NumberOfSnippets - NumberOfUniqueSnippets) / (NumberOfSnippets)

  Inkluderar forkade notebooks, vilket vi inte ska göra!

  Han räknar ut:

  - distributionen av språk i notebooksen. Python dominerar starkt (94,93%). Därefter kommer Julia, R och Scala (0,79%, 0,79% respektive0,16%). Visualiserar med torusdiagram.
  - andel klonade snippets, för varje språk och för hela datamängden. Visualiserar med heat maps.

  Sedan analyserar han vilka paket som importeras tillsammans i Python-notebooks. Räknar ut "normalized level of complementarity" som
  P(A|B) - P(A), där
  P(A|B) = (antal notebooks som innehåller A och B) / (antal notebooks som innehåller A)
  P(A) = (antal instanser av A) / (totalt antal imports)
  P(A) är alltså sannolikheten att en *import* ska vara A, medan P(A|B) snarast verkar vara den betingade sannolikheten för att en *notebook* innehåller B givet att den innehåller A (normalt betecknad P(B|A)).

  Analysen sker för var och en av grupperna
  1) 0-50 kB
  2) 50-100 kB
  3) 1-30 MB

  De vanligaste paketen (i alla grupperna) är numpy, pandas och matplotlib. (Dessa används ofta tillsammans och brukar kallas för "the scientific Python stack".) Ju större filer, desto större sannolikhet för import av var och ett av dessa 3 paket (och även några andra).

  Han kommer vidare fram till att:

  - 12,6% av notebooksen innehåller bara unika snippets
  - Ungefär 75% av alla snippets finns i minst 2 olika notebooks
  - 53,9% av alla notebooks består helt och hållet av kloner
  - en genomsnittlig notebook består till 69,8% av klonade snippets

  Pratar om att alla negativa korrelationer är små, men verkar inte räkna ut några korrelationer mellan någonting. (Förmodligen menar han negativ "normalized level

och complementarity".)