Exploration and Explanation in Computational Notebooks

Typ	Konferensartikel
Författare	Adam Rule
Författare	Aurélien Tabard
Författare	James D. Hollan
Webbadress	http://dl.acm.org/citation.cfm?doid=3173574.3173606
Ort	Montreal QC, Canada
Utgivare	ACM Press
Sidor	1-12
ISBN	978-1-4503-5620-6
Datum	2018
DOI	10.1145/3173574.3173606
Hämtad den	2019-09-26 13:03:17
Bibliotekskatalog	Crossref
Namn på konferens	the 2018 CHI Conference
Språk	en
Sammanfattning	Computational notebooks combine code, visualizations, and text in a single document. Researchers, data analysts, and even journalists are rapidly adopting this new medium. We present three studies of how they are using notebooks to document and share exploratory data analyses. In the first, we analyzed over 1 million computational notebooks on GitHub, finding that one in four had no explanatory text but consisted entirely of visualizations or code. In a second study, we examined over 200 academic computational notebooks, finding that although the vast majority described methods, only a minority discussed reasoning or results. In a third study, we interviewed 15 academic data analysts, finding that most considered computational notebooks personal, exploratory, and messy. Importantly, they typically used other media to share analyses. These studies demonstrate a tension between exploration and explanation in constructing and sharing computational notebooks. We conclude with opportunities to encourage explanation in computational media without hindering exploration.
Protokolltitel	Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18
Tillagd den	2019-09-26 13:03:17
Ändrad den	2019-10-07 12:43:54

Etiketter:

GitHib mining, Jupyter

Bifogade dokument

- Rule m. fl. - 2018 - Exploration and Explanation in Computational Noteb.pdf

  Författarna har gjort 3 delstudier.

  ## Studie 1

  I den första analyserar de de 1 227 573 publika notebooks på github som var nedladdningsbara i juli 2017, och som inte var forkade från ett annat repositorie och kommer fram till att:

  - Antalet notebooks per användare och per repro följer en exponentiell (avtagande) fördelning.
  - Språket fanns specificerat i 85,1% av notebooksen. Bland dessa 85,1% använde:
    - 96,3% Python
    - ca 1% R
    - ca 1% Julia
  - Av notebooksen skrivna i de 3 ovan nämnda språken importerar 89,1% externa paket/moduler. De vanligaste i Python var:
    - numpy (67,3%)
    - matplotlib (52,1%)
    - pandas (43,3%)
  - 99,8% av alla celler var antingen kod- eller markdown-dito
  - 27,6% av notebooksen innehöll ingen text, utan bara bilder och kod
    - Om man bortser från dessa följde både #celler, mängden text och LOC per notebook log-normalfördelningar.
  - 2,2% av notebooksen innehöll ingen kod
  - Medianantalet ord/notebook var 218 om man bortsåg från de som inte innehöll någon text. (Största var 55 000 ord.)
  - Medianantalet kodrader var 85 om man bortsåg från de som inte innehöll någon kod. (Största var >400 000 LOC.)
  - Textceller återfanns oftast i början av notebooksen.
  - De vanligaste orden i reprobeskrivningarna var learning, project, machine, udacity, course, deep, nanodegree, neural, kaggle och model, vilket enligt författarna tyder på att notebooks ofta används för utbildning och maskininlärning.
  - Många av notebooksen innehöll ingen story utan var bara en samling skript med några lösa anteckningar.

  ## Studie 2

I den andra studien har de valt ut 52 repron innehållande 221 notebooks kopplade till forskning (=har arXiv- eller DOI-länk i README. Här såg de att:

- Hälften av repron innehöll bara 1 notebook
- I de två repron som hade flest notebooks var notebooksen ofta kloner av varandra (inom ett repro, inte mellan). Dessa rensades bort innan de påbörjade sin analys.
- De återstående 145 notebooksen var längre än de i första studien,
- 55% hade inledande text, men bara 3% hade en avslutande textcell.
- Av de som innehöll text:
  - beskrev 88% de olika stegen i analysen
  - beskrev 34% resonemangen
  - diskuterade 38% resultaten
- 82% av notebooksen innehöll kodkommentarer, i vilka i princip alla beskrev vad programmet gör, 10% resonemang och 4% resultat. 50% innehöll bortkommenterade kodblock.

De kommer fram till att notebooksen används för iterativa analyser, men mindre ofta "rich narratives".

Studie 3

I den tredje studien intervjuar de 15 akademiska dataanalytiker och kom fram till att notebooks används för olika syften, många inom utbildning. Intervjuerna fokuserar dock på de notebooks som används för forskning.

Notebooksen beskrevs ofta som något man använde för att experimentera. Resultat presenterade man ofta i andra medier.

Det var dock några intervjuobjekt som tyckte att notebooks var bra för interaktion med icke-programerare.

Intervjuobjekten kände ofta att deras notebooks behövde städas, oavsett om de skulle delas eller bara användas av dem själva.

Författarna kommer fram till att det finns en motsättning mellan olika användningsområden för notebooks. Om man använder dem för iterativt experimenterande får man ofta röriga notebooks. Ska man använda notebooksen för att dokumentera eller kommunicera sin forskning behöver man städa upp dem.

Författarna tycks anse att notebooksen ska användas i det senare syftet och föreslår åtgärder för att få dataanalytiker att göra det.

Notera

att de även diskuterar begränsningar i studien! Kan vara värt att titta på när vi skriver våra "threats to validity".