Git for Social Scientists: Introduction to Version Control with Git

Tomoya Sasaki

Massachusetts Institute of Technology

November 4th, 2022

Introduction

- Why Git? Why Github? Why version control?
- These are essential tools for programmers
- How about social social scientists?
- My opinion: Social scientists also benefit from version control with Git
  - Increase in collaborative projects
  - Demand for clean replication materials
  - Complex data manipulation/preprocessing/analysis
- "Code and Data for the Social Sciences: A Practitioner's Guide" by Gentzkow and Shapiro has a chapter dedicated for version control
- This workshop
  - Introduction to version control
  - Pros and cons of Git/Github
  - Brief introduction of these tools

What is version control?

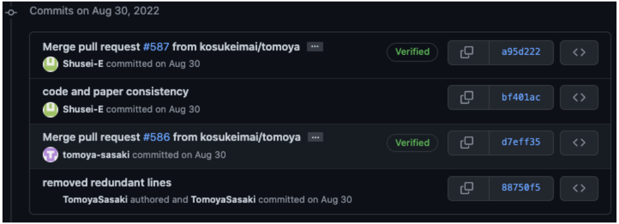
- Version control: tracking and managing changes to file content
- Git: (the most popular) software for version control
- Github: service to host your git on the Internet (alternatives include GitLab, Bitbucket ...)
- Repository: unit of a version control project, contains a folder with a subfolder named `.git`
    - Local repository: repository (folder) in your own computer
    - Remove repository: repository (folder) in a web hosting services such as Github
- `.git` folder in a repository tracks and stores every single change you make in the corresponding repository
- I focus on Git and Github because they are extremely popular than their alternatives

Why Git and Github: Tracking who/how/when

- You can identify
  - who made changes
  - how they made the changes
  - when they made the changes
- You can check the entire history since you created a repository and move back to previous versions easily
- Github can visualize them nicely

- Useful when you
  - want to revert your (particular) changes
  - work on a collaborative project

- You don't need to keep
  - different versions of the same file: `clean_data_1104.R`, `clean_data_1020.R`
  - the same file edited by different people: `clean_data_tomoya.R`, `clean_data_adam.R`



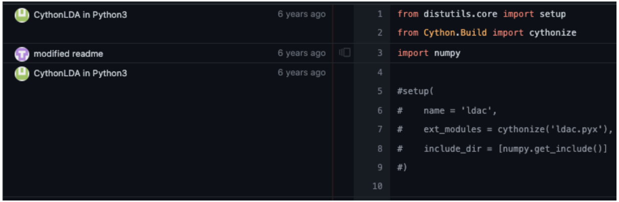
Commits on Aug 30, 2022

Merge pull request #587 from kosukeimai/tomoya ···
Shusei-E committed on Aug 30

code and paper consistency
Shusei-E committed on Aug 30

Merge pull request #586 from kosukeimai/tomoya ···
tomoya-sasaki committed on Aug 30

removed redundant lines
TomoyaSasaki authored and TomoyaSasaki committed on Aug 30



CythonLDA in Python3 — 6 years ago
modified readme — 6 years ago
CythonLDA in Python3 — 6 years ago

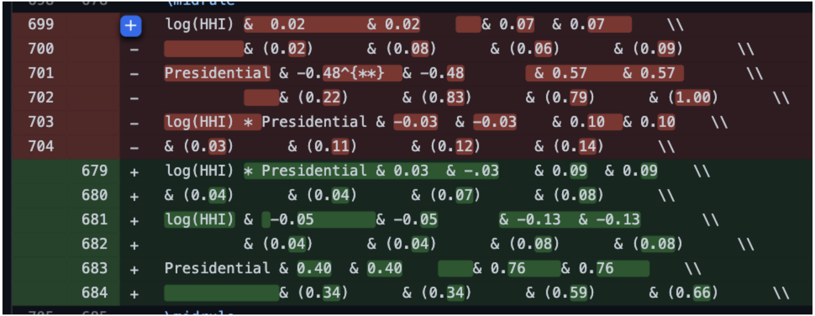
```
from distutils.core import setup
from Cython.Build import cythonize
import numpy

#setup(
#    name = 'ldac',
#    ext_modules = cythonize('ldac.pyx'),
#    include_dir = [numpy.get_include()]
#)

```

Why Git and Github: Tracking who/how/when

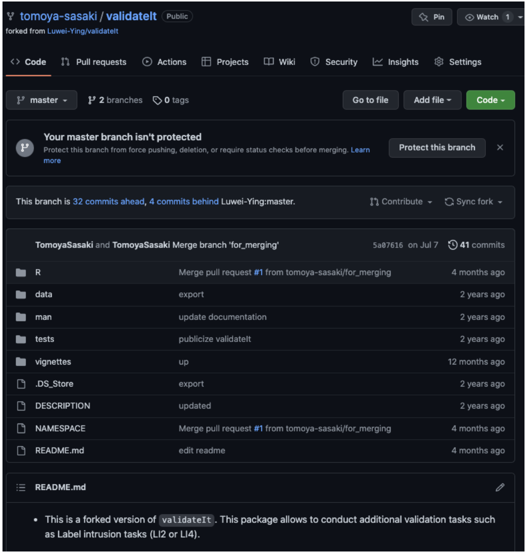
- You can check how the results change when we try different specification
- Easy to track which part of the results changed



```
    \midrule
    log(HHI) &  0.02            & 0.02        & 0.07   & 0.07          \\
-            & (0.02)        & (0.08)        & (0.06)        & (0.09)        \\
-   Presidential & -0.48^{**}  & -0.48          & 0.57        & 0.57           \\
-                & (0.22)        & (0.83)        & (0.79)        & (1.00)        \\
-   log(HHI) * Presidential & -0.03  & -0.03      & 0.10   & 0.10      \\
-   & (0.03)        & (0.11)        & (0.12)        & (0.14)        \\
+   log(HHI) * Presidential & 0.03  & -.03      & 0.09  & 0.09     \\
+   & (0.04)        & (0.04)        & (0.07)        & (0.08)        \\
+   log(HHI) &  -0.05          & -0.05          & -0.13  & -0.13           \\
+            & (0.04)        & (0.04)        & (0.08)        & (0.08)        \\
+   Presidential & 0.40  & 0.40        & 0.76      & 0.76          \\
+                & (0.34)        & (0.34)        & (0.59)        & (0.66)        \\
    \midrule
```

Other side benefits of Git/Github

- Hosting a customizable website (free, no ads, tons of templates)
- Contribute to software packages hosting on Github
- Tweak a package developed by someone else for your own purposes
- Send a request to package developer (often happens at "Issue")
- Nice integration with popular apps/websites such as Rstudio and Overleaf



tomoya-sasaki / validateIt Public

forked from Luwei-Ying/validateIt

2 branches 0 tags

Your master branch isn't protected

Protect this branch from force pushing, deletion, or require status checks before merging. Learn more

This branch is 32 commits ahead, 4 commits behind Luwei-Ying:master.

TomoyaSasaki and TomoyaSasaki Merge branch 'for_merging' 5a07616 on Jul 7 41 commits

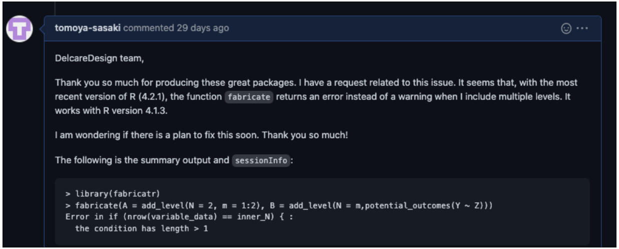
R	Merge pull request #1 from tomoya-sasaki/for_merging	4 months ago
data	export	2 years ago
man	update documentation	2 years ago
tests	publicize validateIt	2 years ago
vignettes	up	12 months ago
.DS_Store	export	2 years ago
DESCRIPTION	updated	2 years ago
NAMESPACE	Merge pull request #1 from tomoya-sasaki/for_merging	4 months ago
README.md	edit readme	4 months ago

README.md

- This is a forked version of `validateIt`. This package allows to conduct additional validation tasks such as Label intrusion tasks (LI2 or LI4).

Other side benefits of Git/Github

- Hosting a customizable website (free, no ads, tons of templates)
- Contribute to software packages hosting on Github
- Tweak a package developed by someone else for your own purposes
- Send a request to package developer (often happens at "Issue")
- Nice integration with popular apps/websites such as Rstudio and Overleaf



tomoya-sasaki commented 29 days ago

DelcareDesign team,

Thank you so much for producing these great packages. I have a request related to this issue. It seems that, with the most recent version of R (4.2.1), the function `fabricate` returns an error instead of a warning when I include multiple levels. It works with R version 4.1.3.

I am wondering if there is a plan to fix this soon. Thank you so much!

The following is the summary output and `sessionInfo`:

```
> library(fabricatr)
> fabricate(A = add_level(N = 2, m = 1:2), B = add_level(N = m,potential_outcomes(Y ~ Z)))
Error in if (nrow(variable_data) == inner_N) { :
  the condition has length > 1
```

Limitations: not suitable to track large files

- Github imposes file size limits:
  - 25 MB per file limit (you can change this limit up to 100MB by changing setup)
  - 1GB per repository limit
- Remember that `.git` tracks and stores all the change you make in a repository

  ↝ if you store a huge file in the repository and let `.git` tracks its changes, the `.git` folder can grow quite huge
- Use `.gitignore` to specify files that Git should ignore

  `.gitignore`

  ```
  *.csv # ignore csv files
  /data/ # ignore data folder
  ```
- Include huge files as well as sensitive files (password, API key etc) in `.gitignore`

Limitations: difficult to track non-text files

-