Lecture 6: more pandas (and git/GitHub)

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- git and GitHub: Let's be more disciplined!
- Python's pandas library

- Tools:
  - Git and GitHub
  - Jupyter Notebook
  - Markdown

Pull trouble? Let's shoot.

- You are trying to pull from upstream, and git says conflicts! Fatal, even!

← Chances are your local file is one that needs to be discarded.

- Pay attention to the conflict message. Which file is causing trouble?
- Copy that local file into a location outside of the repo, just in case
- Then, do:
  - `git rm troublefile`
  - `git commit -m "trouble making file gone"`
  - `git pull upstream master` (← this step might not even be necessary)
- After that, if needed, put back your original file with a different name

Also: keep checking
`git status`
in between

- Still stuck? Google. Look up stack overflow. Email Na-Rae with a screenshot.

Some house rules for happy collabo-gitting

- To add files for committing, use individual file names.
  - Use: `git add file1 file2`
  - Do not add and commit unnecessary files/dirs!

Stop using:
```
git add *
git add .
```

- To delete previously committed files, do ***git-delete***, not just delete.
  - `rm filename` (or drag file to trash bin), immediately followed by `git rm filename`
  - or just do `git rm filename`

`git rm` tells git to stop tracking the file. It also deletes the file itself if it still exists.

- Keep files small (especially for shared repos)
  - GitHub has 100MB file size limit.
  - In general, keep your data files < 3MB.

Many ways of git add/rm

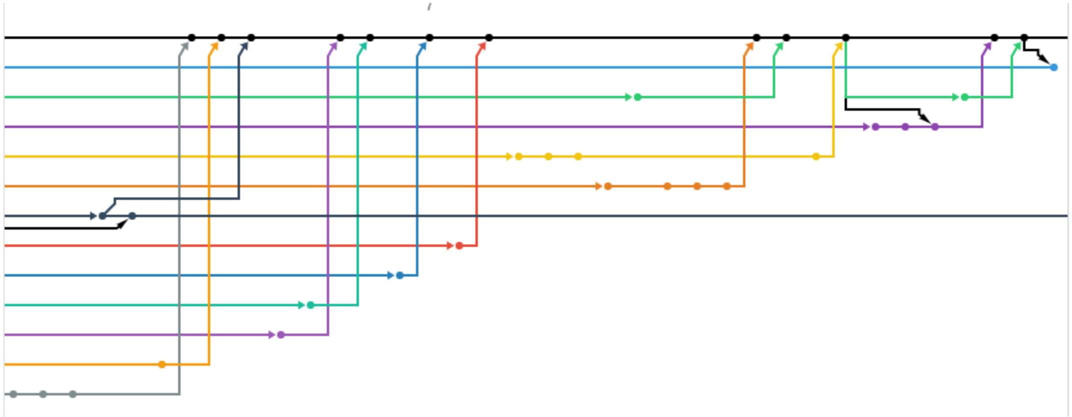
- Different calls have different behaviors.

	New files	Modified files	Deleted files	Files beginning with .	Notes
`git add file1 file2`	✓	✓	✗	✓	Use TAB completion
`git add *`	✓	✓	✗	✗	**AVOID!** Only use like git add *.txt
`git add .`	✓	✓	✓	✓	Any changes in the directory. **AVOID!**
`git add -A`	✓	✓	✓	✓	Likewise. **AVOID!**
`git add -u`	✗	✓	✓	✓	Any updates.
`git rm file1`	If file1 exists, delete and stop tracking.		✓	✓	Opposite of 'git add'. Use with deleted files or to delete files.

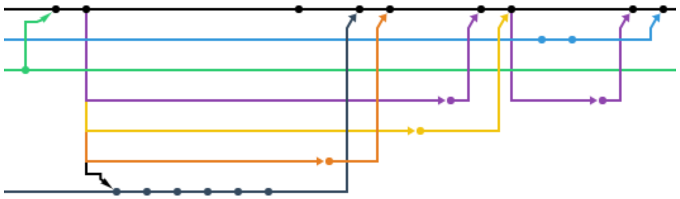
* Files and directories work the same.

Class-Practice-Repo: we'll keep it going

- Let's not resort to "scrap fork/repo and start fresh" again.
- Keeping to a few git/GitHub house rules will hopefully keep our Class-Practice-Repo relatively complication-free.
- Before:



- Now:



Lost continuity. ☹

git is better in color (actually, everything is)

- Windows folks are using git bash, which has nice colorized git output
- Mac users: there are ways to customize your Terminal and git
  - Brianna will demonstrate her setup

Dealing with a Private repository

- We now have some **private** repos:
  - HW2-Repo, Licensed-Data-Sets
- If you get an error while cloning them, you might need to alter your clone URL.
  - This is the usual one (copied through GitHub's green button):

    ```
    git clone https://github.com/NAME/REPO.git
    ```

  - If that fails, try this one instead:

    ```
    git clone https://username@github.com/NAME/REPO.git
    ```

    🡐 Forces fresh credential check

    🡐 Changed1: you should put your actual user name string in the command line

    🡐 Changed2: I originally had password in the line as well, but apparently that is ill-advised (will be recorded in git history). Leaving it out makes git prompt for your password, which is better.

pandas practice, continued

Activity 3
50 mins

- 50 Years of Pop Music
  - http://kaylinwalker.com/50-years-of-pop-music/
  - Download CSV file 'billboard_lyrics_1964-2015.csv'

- In Class-Practice-Repo, activity3 folder:
  - Move or copy the CSV file into the directory.
  - You will find `pop_music_lyrics.ipynb`
  - ← Rename it `pop_music_lyrics_YOURNAME.ipynb` and work on it.

Homework 2: Process ETS Corpus

- http://www.pitt.edu/~naraehan/ling1340/hw2.html
- Corpus distributed via private GitHub repo "Licensed-Data-Sets"
  - https://github.com/Data-Science-for-Linguists/Licensed-Data-Sets
  - No need to fork: clone directly. (Why?)
  - If you have trouble cloning, see slide #8.

Wrapping up

- To-do 5: due Tuesday.
  - Ultimate pandas notebook. Also: visualization.
- HW2: Process ETS Corpus
  - Due next ~~Tuesday~~ → THURSDAY
  - Don't wait -- get started this weekend! This HW is in period!
- Project ideas: you should start thinking.
- Learn:
  - pandas
  - matplotlib, visualization