Lecture 5: pandas

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Python's pandas library
- Tidying up GitHub's Class-Practice-Repo
- Notes on assignment grading, learning from assignments

- Tools:
  - Git and GitHub
  - Jupyter Notebook

Back to Class-Practice-Repo

https://github.com/Data-Science-for-Linguists/Class-Practice-Repo

Your .gitignore file should look like this:

```
narae@T450s MINGW64 ~/Documents/Data_Science/Class-Practice-Repo (master)
$ cat .gitignore
# ignore directories named 'foo' anywhere
**/foo

# ignore jupyter notebook checkpoint directories anywhere
**/.ipynb_checkpoints
```

- If not, you should replace your file with the correct file.

1. Remove your file:

```
rm .gitignore
```

2. Create a new file via text editor. Editor window opens up.

```
atom .gitignore
```

3. Copy over the correct .gitignore content from GitHub, paste, and save.

Notes about homework

- Policies in http://www.pitt.edu/~naraehan/ling1340/policies.html
- To-Dos are completion-based.
  - 0 for no work, 5 for unsatisfactory (<70%) work, 10 for satisfactory work
  - I will likely not provide detailed feedback, if at all.
- Homework
  - Pay attention to instructions.
  - I will be posting grade and feedback on CourseWeb.
  - You should learn from your classmates' submissions.
  - Correctly working code is only part of homework objectives.
  - Communicative and presentational aspects are as important.
  - Treat your Jupyter Notebook as a written report with bits of code embedded.

To-Do 3

- What exciting spreadsheets did you all submit?

🡐 Next To-Do: try a spreadsheet submitted by a classmate

pandas practice

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-Dat-Sets"
  - https://github.com/Data-Science-for-Linguists/Licensed-Data-Sets

Wrapping up

- To-do 4: due Thursday.
  - Continuation of To-do 3. Submit through todo3/ directory in Class-Practice-Repo.
- HW2: Process ETS Corpus
  - Due next ~~Tuesday~~ → THURSDAY
  - You should get started!
- Project ideas: you should start thinking.
- Learn:
  - pandas
  - Git, GitHub
  - Jupyter Notebook