Lecture 8: Corpus Projects, Data Formats

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Homework 2 wrap up
- Corpus linguistics
  - Gries & Newman (2013) "Creating and using corpora"
- Tools:
  - Corpora and corpus tools
- Your term projects
  - Data management: copyright and licensing
- Data formats

Homework 2 ETS corpus continued

- Jupyter notebook files in Na-Rae's directory:
  - https://github.com/Data-Science-for-Linguists/HW2-Repo/tree/master/narae
- PART 1
  - Processing CSV files, test/train/development split
  - Graphs
- PART 2
  - Group-by, aggregation, stack/unstack, pivoting...
  - "Titanic" example data set
  - Configuring visualization style
- PART 3
  - Processing response files and prompt files
  - Tokenization, computing various stats

Corpus linguistics

- Review Gries & Newman (2013) "Creating and using corpora"
- Review "A list of corpora and corpus resources"
  - https://github.com/Data-Science-for-Linguists/Corpus-Resources/blob/master/corpus_tools_list.md

Your term project

- Your project is now on GitHub
  - https://github.com/Data-Science-for-Linguists
- First progress report next Thursday
  - Focus on data: sourcing, curation and cleaning
  - Some of you might discover your plan is not feasible.
    - You will need to find a new project idea and plan quickly!
- Managing your data
  - You will be manipulating and processing your data.
  - Read: The Basic Reproducible Workflow Template
  - Keep a few versions along with the code that produced them.
  - Should you include your data set in your GitHub repo?

    GOOD QUESTION. Next slide ➔

Licensing, public vs. private

- Your data:
  - Your original data source: what kind of license does it come with?
  - Can you re-distribute the data?
  - "Derivative" data: are you allowed to distribute?
  - How about samples?
  - How to best *present* the outcome and ensure *reproducibility* if you cannot share your data in full?
- Your code:
  - Will you allow other people to use your code? Re-distribute?
  - Will you allow other people to turn your code into a commercial product? Patent it?

Licensing, public vs. private

- As a principle, your term project -- including code and data -- should be **as public and open as possible**.
  - Set your repo to **public** at this time.
    - Justification needed for changing to private.
  - For now, store your data files in data/ directory, and have git ignore this directory through `.gitignore` file, like below:

```
narae@T450s MINGW64 ~/Documents/Data_Science/Inaugural-Address-Project (master)
$ ls
LICENSE.md  README.md  data/

narae@T450s MINGW64 ~/Documents/Data_Science/Inaugural-Address-Project (master)
$ echo "**/data" > .gitignore

narae@T450s MINGW64 ~/Documents/Data_Science/Inaugural-Address-Project (master)
$ ls -la
total 19
drwxr-xr-x 1 narae 197121   0 Oct  3 17:56 ./
drwxr-xr-x 1 narae 197121   0 Sep 26 16:08 ../
drwxr-xr-x 1 narae 197121   0 Aug 28 16:32 .git/
-rw-r--r-- 1 narae 197121   8 Oct  3 17:56 .gitignore
-rw-r--r-- 1 narae 197121 385 Jul 28 16:13 LICENSE.md
-rw-r--r-- 1 narae 197121 120 Jul 28 17:00 README.md
drwxr-xr-x 1 narae 197121   0 Aug 28 16:32 data/

narae@T450s MINGW64 ~/Documents/Data_Science/Inaugural-Address-Project (master)
$ cat .gitignore
**/data
```

Licensing, public vs. private

- Do your research on copyright and licensing.
  - http://www.library.pitt.edu/copyright
  - https://choosealicense.com/
- Document, document, document!
  - You should **document and justify** your sharing and licensing decisions. It is an important part of your project.

Data standards & exchange formats

	What	Notes, reference
CSV	Comma-separated values	Compatible with Excel
TSV	Tab-separated values	
HTML	Web pages	
XML	For markup and text encoding	A Gentle Introduction to XML by TEI
JSON	JavaScript Object Notation Twitter, Jupyter Notebook	Introducing JSON JSON example (vs. XML)

They are all TEXT files.

- Encoding: Latin-1, ASCII, UTF-8, UTF-16, CP1252, ...
- Line endings:
  - LF (`'\n'`: OS X & Linux) , CRLF (`'\r\n'`: Windows)
- But underneath it all, these files are all TEXT files with **special formatting syntax** and **special characters** designated for formatting purposes.
  - In command line, you can `cat` and `less` through the files.
  - You can open them up in a text editor (Atom, Notepad++) and edit.
  - Some editors/applications are aware of the format-specific syntax and will highlight/render accordingly.
    - Unlike, say, PDF files, style attributes are NOT part of the files themselves. (e.g., markdown file)

Do not re-invent the wheel.

- Don't try and parse them manually.
- There are Python libraries. Import and use them.
  - CSV & TSV: `pandas`
  - HTML & XML: Beautiful Soup (`bs4`)
  - JSON:
    - `json` library
    - `pandas.read_json`

Project-specific (ad-hoc) formats

- Brown corpus

```
The/at Fulton/np-tl County/nn-tl Grand/jj-tl Jury/nn-tl said/vbd
Friday/nr an/at investigation/nn of/in Atlanta's/np$ recent/jj
primary/nn election/nn produced/vbd ``/`` no/at evidence/nn ''/''
that/cs any/dti irregularities/nns took/vbd place/nn ./.
```

- Korean Treebank corpus:

```
;;05:127: 저는 그 일을 할 수 있는 한 빨리 하겠습니다 .

(S (NP-SBJ 저/NPN+는/PAU)
   (VP (NP-OBJ-LV 그/DAN
                  일/NNC+을/PCA)
       (VP (NP-ADV (S (NP-SBJ (S (NP-SBJ *pro*)
                                 (VP 하/VV+ㄹ/EAN))
                              (NP 수/NNX))
                          (ADJP 있/VJ+는/EAN))
                      (NP 한/NNX))
           (ADVP 빨리/ADV)
           (VP (LV 하/VV+겠/EPF+습니다/EFN))))
   ./SFN)
```

It is up to end users to write code to parse data files.

Refer to documentation!

Format conversion

- When dealing with corpora, you may need to convert 100+ files at once.
  - On-line services are too cumbersome.
  - Try batch-processing through command line.
- Automatic tools available on command line.
  - Encoding conversion: `iconv` (Linux, OS X, on Git Bash)
  - Line ending conversion: `unix2dos`, `dos2unix`
  - Pandoc http://www.pandoc.org/
    - Universal document coverter
    - HTML, XML, PDF, LaTeX, Markdown, Epub, MS Doc, ...

Data-mining web & social media

- Twitter sample corpus
  - Static corpus: download from the NLTK data page
- How does one data-mine Twitter?
  - Answer: through **API** (Application Program Interface)
  - To-do #8
  - Getting acquainted with JSON format
  - Data Analysis using Twitter API and Python, The Code Way tutorial
  - And a couple more on the Learning Resource page

Wrapping up

- To-do 8 due on Thursday.
- Read for class discussion:
  - The Basic Reproducible Workflow Template
  - Think about licensing issues for your project
- Reminder:
  - REFRESH your browser window.
  - Class schedule page & Learning resource page are frequently getting updated.