Lecture 8: Corpora and Data Formats

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Your term project
  - Plan submitted, repo created!
  - Work on your DATA!
- Corpus data: standard and popular formats
  - Encoding, line break
  - Review of common data formats
- Homework 2 review wrap

Your term project

- Everyone's project repo is at our GitHub org.
- First progress report is due next Friday!
  - Focus on data: sourcing, curation and cleaning
- Managing your data
  - You will be manipulating and processing your data.
  - Should you include your data set in your GitHub repo?
    - Depends!

Data standards & exchange formats

	What	Notes, reference
CSV	Comma-separated values	Compatible with Excel
TSV	Tab-separated values	
HTML	Web pages	Not meant as data format
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)

These are all TEXT files!

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. Also: `head`, `tail`
  - 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)

Wrapping up

- No To-do out
  - Work on your project!
- Your project
  - Work on it! Focus on DATA.