Lecture 7: Corpus Linguistics

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Homework 2 continued
- Corpus linguistics
  - Gries & Newman (2013) "Creating and using corpora"
- Tools:
  - Corpora and corpus tools

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

Another licensed data set

- TIMIT Acoustic-Phonetic Continuous Speech Corpus
  - https://catalog.ldc.upenn.edu/ldc93s1
  - In "Licensed-Data-Sets" repo
  - Is this a "corpus"...?

Your project ideas → project plan

- Most of you received feedback on your project ideas.
- Project plan due on Tuesday.
    - http://www.pitt.edu/~naraehan/ling1340/project.html#plan
    - Create a GitHub project repo.
    - Incorporate my feedback!

Wrapping up

- Project plan due Tuesday.
  - http://www.pitt.edu/~naraehan/ling1340/project.html#plan

- No other assignment due. You should:
  - Catch up on pandas and visualization.
  - Re-visit Homework 2.
  - Learn more about git/GitHub.
  - Start data sourcing portion of your project.
  - Take a look at classmates' projects.