ANLY 520: Final Project

Erin M. Buchanan

2019-01-08

Description

The term project for this class, due the last week of the semester, consists of a presentation demonstrating the use of the Natural Language Processing techniques you've learned in the course. You will choose a data set, and use the tools of your choice to analyze the data. You will create a report of the project in *Jupyter* and present your findings. The code for the project should be inline with the text, although, you may like to create a separate script to first start the project. You will turn in a link to an audio presentation of your report (5-10 mins), along with the report.

Groups

You may work in small groups of two-three people if you have the same idea for a project. You will need to coordinate the video presentation, such that each person presents a portion of the project.

Data

You may choose any data set you would like to work on, as long as it contains at least 1000 distinct unstructured texts. This data can be a collection of Twitter data, blog posts, e-mails, news reports, or similar data. The following links contain some suggested data sets:

- Enron Email dataset: https://www.cs.cmu.edu/~./enron/
- Stanford Sentiment Analysis dataset (Twitter): http://nlp.stanford.edu/sentiment/
- FBS opinion mining data sets: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets
- Cornell movie review data sets: https://www.cs.cornell.edu/people/pabo/movie-review-data/

Processing

Once you've chosen a dataset, you will perform the following tasks:

- Clean the data for processing
- Create a corpus appropriate for your tool
- Tokenize the data
- Perform basic analysis of the data
- Train a sentiment analytic
- Test the sentiment analytic

Reporting

Your report should consist of the following items, supported by appropriate text, diagrams, and code:

- A description of the dataset, including where it was acquired from, the format of the data, and whether or not the data included any tagging.
- Basic statistics about the data
  - Number of documents
  - Average length of document (characters and words)

  - Frequency distribution/lexical dispersion
- Explanation of your sentiment analytic tool and method, including any source code you produce
- Explanation of your training and test set creation
- Results of testing your analytic
  - Findings of the analytic
  - Precision
  - Recall
  - F-score
  - Performance (average processing time per document)
- Your conclusions, including anything notable you learned in the process of completing the project.
- Remember to properly document your sources.

What to Turn In

- A link to a video of your presentation:
  - Windows or Mac: You can use loom.com and the Chrome browser extension to record your video (free!).
  - Mac: You can use QuickTime to record your screen (the free version that comes with Macs).
  - Share the link to the loom video or upload the video to YouTube/other sharing source.
- A Jupyter notebook of your presentation:
  - You can compile the notebook into PDF/HTML for easier comments/grading.