Text as Data: Homework 3

Assigned: 5/9, Due 5/16

In this homework we will analyze a collection of news stories from the New York Times from the November 1-3, 2004 (the day before, of, and after the 2004 general election). This data come from the New York Times Annotated Corpus and is for academic use only. We have done some preprocessing in order to simplify the homework tasks.

1 Preprocessing and Creating a Document-Term Matrix

a) From the course github, download `nyt_ac.json`

b) Using the `JSON` library in python, import the data. Use `type` to explore the structure of this data. How are this data organized?

c) Extract the title and text from each story. Create an individual document for each story and write each of the files to a new directory

d) Using the loaded `json` file, create a document term matrix of the 1000 most used terms. Be sure to:

- Discard word order
- Remove stop words
- Apply the porter stemmer

e) Include in your document-term matrix the *desk* from which the story originated, which we will include later

2 Dictionary Classification Methods

a) Download the list of positive

https://raw.githubusercontent.com/nealcaren/quant-text-fall-2014/master/positive.csv and negative

https://raw.githubusercontent.com/nealcaren/quant-text-fall-2014/master/negative.csv words from Neil Caren's website.

b) Calculate a positive score and a negative score for each document and the difference between each score using the dictionaries

c) How does the score change before and after the election? How does the score vary across desks?

3 Supervised Learning with Naive Bayes

a) Let's focus on documents that came from Business/Financial desk and National Desk. Using leave-one out cross validation, calculate the accuracy of Ridge Regression to calculate the label.

b) Compare the performance of Ridge to the performance of 2 of the following 3 algorithms using 10-fold cross validation:

- LASSO
- Elastic Net, $\alpha = 0.5$
- Elastic Net, $\alpha = 0.25$

How does Ridge compare?