Text as Data: Homework 3

August 17, 2017

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

Clustering Methods

1) Using the `kmeans` function, create a plot of the `kmeans` objective function as the number of clusters varies from 2 to $N - 1$.

2) Apply K-Means with 6 clusters, being sure to use `set.seed` to ensure you can replicate your analysis

3) Label each cluster using computer and hand methods:

   i) Suppose $\boldsymbol{\theta}_k$ is the cluster center for cluster k and define $\bar{\boldsymbol{\theta}}_{-k} = \frac{\sum_{j \neq k} \boldsymbol{\theta_j}}{K-1}$ or the average of the centers not k. Define

   $$\text{Diff}_k = \boldsymbol{\theta}_k - \bar{\boldsymbol{\theta}}_{-k}$$

   Use the top ten words from Diff_k to label the clusters

   ii) Sample and read texts assigned to each cluster and produce a hand label

2 Dictionary Classification Methods

a) Download the list of positive (http://www.unc.edu/ ncaren/haphazard/positive.txt) and negative (http://www.unc.edu/ ncaren/haphazard/negative.txt) stop 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) Using the version of Naive Bayes outlined on slide 24 of lecture 14, write a function to estimate $p(C_k)$ and $\boldsymbol{\theta}_k$ for an arbitrary collection of categories. Hint: to compute the probability of a document from a category, note you can work with the log of the probability equivalently.

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

c) Compare the performance of Naive Bayes to the performance of 2 of the following 3 algorithms using 10-fold cross validation:

   - LASSO
   - Ridge
   - KRLS

   How does Naive Bayes compare?