Text as Data: Homework 5

Due: 12/5 at 5pm

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 website, 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 (we will use this later to run `Mallet`)

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 Using `MALLET` to Fit Topic Models

a) Go to `http://mallet.cs.umass.edu/download.php` and install `MALLET` on your computer

b) Following the syntax from the code posted on the course website (flake.mallet) and from the `MALLET` website, apply LDA with 4 and 8 topics

c) Using the code from the course website `Examp_tc11.R` move the output of `MALLET` into `R` and create a table describe both versions of LDA

d) Compare the 4 and 8 topic models. How do the 4 and 8 topic models differ? What information is included in the 8 topic model that isn't in the 4 topic model

e) For each originating desk, calculate the average proportion of documents from a desk dedicate to each topic. How does the topic attention differ across desks?

3 Using the `Structural Topic Model` in `R`

a) Download the `stm` package for `R` from `CRAN`

b) Convert the document-term matrix to the appropriate format. To do this, create a list in `R` where each component of the list corresponds to an individual document. Store in each component of the list a two rom matrix. The number of columns corresponds to the number of non-zero entries for the document in the document-term matrix. The first row will describe the words used in the document (the columns with the non-zero entry). The second row will correspond to a count of each of the words in the document (they should all be non-zero)

c) Following the help file in `STM` fit a model with 8 topics that conditions on the `desk` of origin for topic prevalence

d) Use `labelTopics` to label each of the topics

e) Compare the 8 topic proportions for each document to the 8 topic proportions without conditioning on `desk` (in vanilla LDA). How do the results differ?

4 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?