Text as Data: Homework 3

Assigned 1/31, Due 2/7

In this homework assignment we're going to continue comparing the press releases of two senators—Richard Shelby and Jeff Sessions, Republican senators from Alabama. To make this comparison, we're going to use word separating algorithms, topic models, and principal components. To make all of these comparisons you're going to use the DTM from the last homework. It should have the press releases on the rows and the set of unigrams and trigrams on the columns.

Applying Word Separating Algorithms

1) Using the document-term matrix, for both unigrams and trigrams create the following three measures of word separation

   i) Independent linear discriminant$\leadsto$ measure used in Mosteller and Wallace (1963)

   ii) Standardized mean difference$\leadsto$ For each word J calculate:

   $$\text{std diff} \quad = \quad \frac{\text{Difference in author means}}{\text{Standard error, diff. in means}}$$

   iii) Standardized Log Odds$\leadsto$ as described in Monroe, Colaresi, and Quinn (2009). To create the scores, set $\alpha_j = 1$

2) Create a plot for each of the measures that shows the most discriminating words. Some helpful functions are
   `plot`, but set `pch = ‘’`
   `text` allows the placement of texts on plots.
   Can we learn anything about how Jeff Sessions and Richard Shelby present their work to their constituents?

3) Compare the discriminating measures in 3 plots. What are the primary differences across the measures?

Low Dimensional Embeddings with Principal Components

1) Wise Will (WW), your friend with a weird name, notices you looking at the slides about principal component analysis (PCA). WW casually remarks that the variance of the eigenvalues of the variance-covariance matrix is a useful heuristic for knowing if PCA can be fruitfully applied to some document-term matrix. WW, completely unsolicited, explains that as the variance of the eigenvalues goes up, the more useful PCA will be. He then laughs and leaves your office. WW is kind of a jerk.

   Let's formalize WW's suggestion. Suppose document-term matrix $\boldsymbol{X}$ has variance-covariance matrix $\boldsymbol{\Sigma} = \frac{\boldsymbol{X}^{'}\boldsymbol{X}}{N}$. And suppose that $\boldsymbol{\Sigma}$ has eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_d > 0$. Then we calculate the variance of the eigenvalues as

   $$\sigma^2 \ = \ \frac{1}{d}\sum_{j=1}^{d}(\lambda_j - \bar{\lambda})^2$$

   where $\bar{\lambda}$ is $\frac{1}{d}\sum_{i=1}^{d}\lambda_i$. WW is saying that as σ^2 gets bigger, a low-dimensional embedding via PCA will provide a better summary of our data.

   Does WW have a good point? Why or why not? (Hint: what do the eigenvalues represent?)

2) Apply the function `prcomp` to $\boldsymbol{X}^{*}$. Be sure to set use a scaled version of the data, by setting `scale = T`, which will ensure that each column has unit variance.

   a) Create a plot of variance explained by each additional principal component. What does this plot suggest about the number of components to include?

   b) Plot the two-dimensional embedding of the text documents. Label the press releases with the file name.

   c) Label the two largest principal components. What does this embedding suggest about the primary variation in the press releases (Hint: if your `embed` is your object with principal components, examine `embed$rotation`)

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 row 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

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

e) Compute the average proportion in each topic for the Shelby press releases and the average proportion in each topic for the Sessions press releases. Compare the averages, what do you notice?