Content analysis: Worksheet 2

Trinity Term 2015

1. Access a text version of the *Federalist Papers* at `http://thomas.loc.gov/home/histdox/fedpaper.txt` and convert it into two term-document matrices (TDMs). In both of them, each of the Federalist Papers should be one document; punctuation should be removed and all characters should be converted to lower-case. One TDM should include *only* stop words; the other should *exclude* stop words and be stemmed.

   Notes:

   - Observe patterns in the text file that will allow you to efficiently divide the text into separate documents.
   - Note that there is metadata before each document, including the title (e.g. "FEDERALIST No. 49") and purported author (e.g. "HAMILTON", "HAMILTON OR MADISON"). Exclude these from the TDMs but retain them for analysis, perhaps in a separate matrix.

   If you are unable to do this (or uninterested in it), you can access completed TDMs at `http://andy.egge.rs/teaching/content_analysis/fed_papers/`.

   - `StopAllWords.txt` is the TDM of stop words;
   - `NoStopStemAllWords.txt` is the TDM of non-stop words.

   They are tab-delimited files that can be loaded into `R` with `read.delim(filename, row.names = T)`.

2. Use `hclust()` or a similar clustering algorithm to examine similarity among texts ascribed to different authors. Which TDM and what distance/similarity measures capture features that are more predictive of authorship?

   Notes:

   - You must pass a "distance" object to `hclust`. One way to get this object is to apply the `dist()` function directly to the TDM; this yields a matrix of Euclidean distances between each pair of columns. Another way is to (write a function to) calculate the

cosine similarity between each pair of columns to create a symmetric $D \times D$ matrix (let's call it DD) and then pass `as.dist(DD)` to the function. Compare these approaches.

- To get the dendrogram from lecture, `plot(hclust(as.dist(DD), method = "ward.D))`, where DD is the cosine similarity-based matrix derived from the stop words TDM. The labels at each leaf of the plot are taken from the row/column names of the distance matrix you pass, so you might want to use the authors' names and/or the number of each Paper.

3. What else can you do with this data?

   - Try implementing a Naive Bayes classifier or RandomForests classifier using the stop words TDM to label the documents with ambiguous ownership.
   - Does a topic model based on the non-stop-words TDM of the *Federalist Papers* reflect the content divisions as described by the authors?