Homework 4

STAT-GB.4310: Statistics for Social Data
Instructor: Patrick O. Perry

Due March 1, 2016

Application

Replicate the analysis from the February 23 lecture on matrix decompositions, using the `federalist.json` corpus instead of the `classic3.json` corpus. That is, perform the following actions:

1. Represent the 85 federalist papers as a document-by-term matrix. You will need to choose whether to perform stemming and whether to remove stopwords, numbers, and punctuation. You will also need to choose what weighting to use (`weightTf`, `weightTfIdf`, etc.). If you'd like, you can use a POS tagger to help to filter words before constructing the matrix.

2. Compute a rank-k singular value decomposition of the document-by-term matrix, for a suitable choice of k. Justify your choice of k.

3. Use the first two left singular vectors to visualize the 85 documents. Use a different color or plotting symbol for each document.

4. For each of the first 5 right singular vectors, report the 10 terms with the highest loadings. Can you identify "topics" associated with these vectors?

5. Apply k-means to the left singular vectors, scaled by the singular values, to cluster the 85 federalist papers. Tell the `kmeans` function to use 3 clusters, or more if you think it makes sense. Report the agreement between the clusters found by your analysis and the true document authors, using the `table` function.

6. What are the advantages of the analysis here over the analysis performed by Mosteller and Wallace? What are the disadvantages? Explain.