New York Times Articles about “Immigration” Analysis

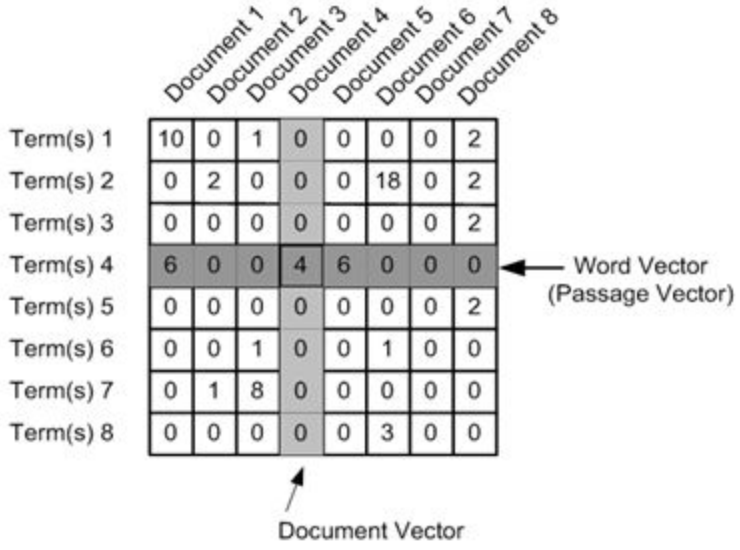
Zhecan Wang

Overview

Based on the New York Times API, I make a script to stream all the articles that mention the keyword ”Immigration” from 1850 to 2005. As a result, I collect around 30000 articles totally. However, in order to better compare this data set with other bills and budget data in time series, I would focus on the articles in the range of 1900 to 2005 (around 15000 totally).

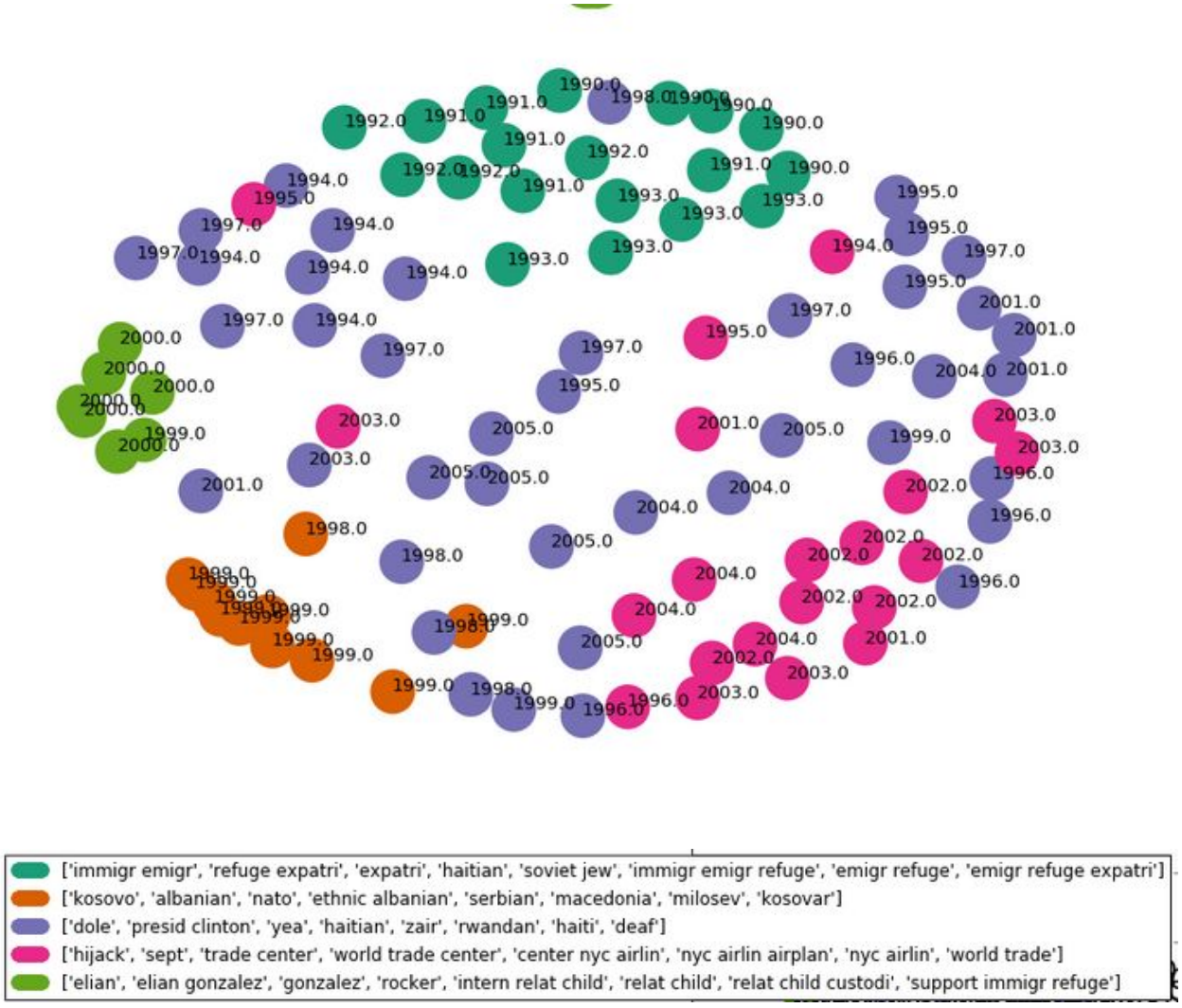
Clustering

I use TF-IDF to analyze all the articles and extract a feature matrix from the data set. In this matrix, each row represents either a group of articles or an article and each column is a feature word. After feeding this matrix in the K-means clustering algorithm, we end up with 5 clusters.



In the following picture, I group the articles together into 100 bins among the 15000 articles, so each circle represents around 150 articles. Each circle is originally a feature vector and should be graphed in a high dimensional space but here I project (smash) them all into a two dimensional plane by using the **Multidimensional Scaling** (reduction).

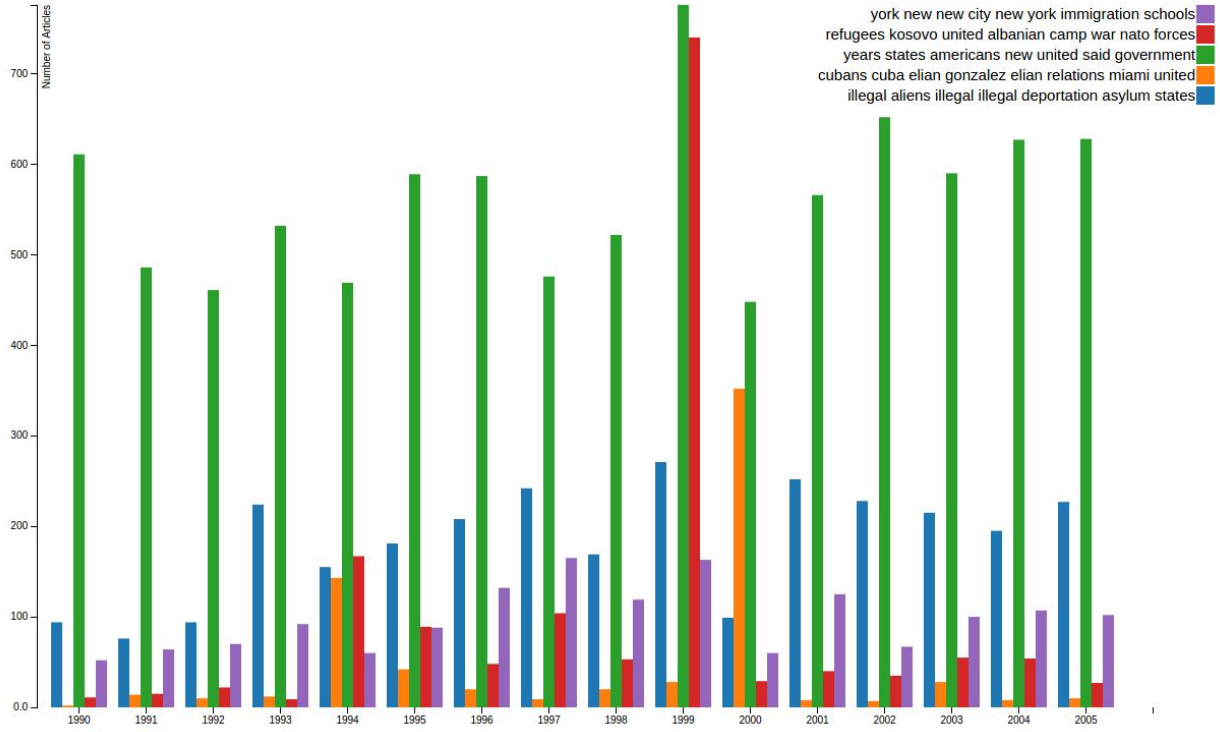
As you may observe, the circles in a cluster tend to be closer together in spatial distance. Also, the year of circles in a cluster also tend to be closer in chronological distance. The clusters' keywords are very distinct from each other. Overall, it is a reasonable clustering.



The following is the overall clusters distribution across years.

Among the five clusters, there are two which may be more interesting. The first one is the "kosovo" (second in order) cluster that tends to center around 1999. The background context can be referred to the Kosovo War, Kosovo Albanians and so on. The second one is the "cubans" (fourth) cluster that tends to center around 2000. The background context can be referred to the Elian Gonzalez custody battle, former Cuban

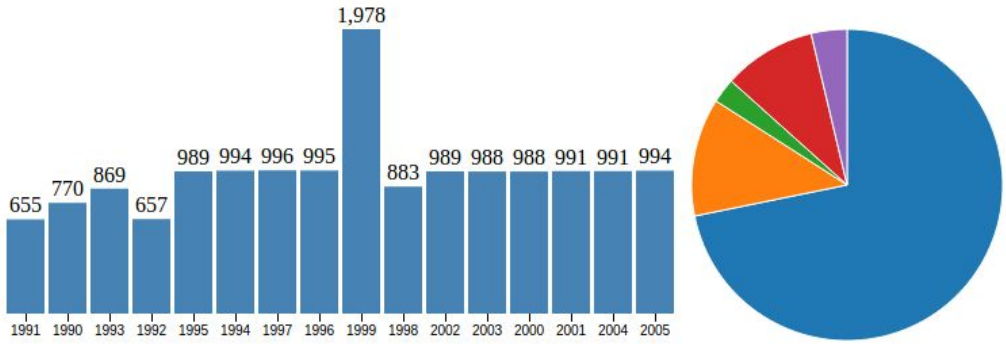
castaway Elián González and so on. The other clusters tend to be more general comparatively.



After clustering, we grab the center words from each cluster along with their importance score (from TF-IDF) and use word cloud to visualize. The following picture shows the important words in the “kosovo” (second in order) cluster. The interesting words here are “kosovo”, “albanian”, “serbian” and so on.

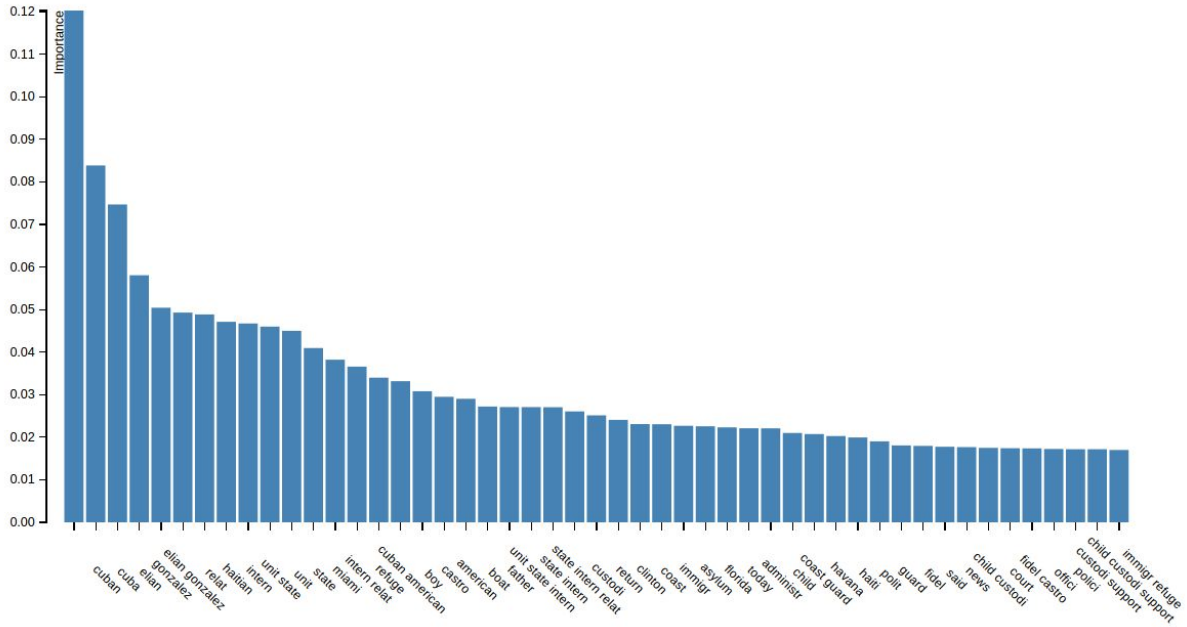
Visualization

The following is another D3 visualization I created for the clustering year distribution (The actual interactive script can be found in the github repository).

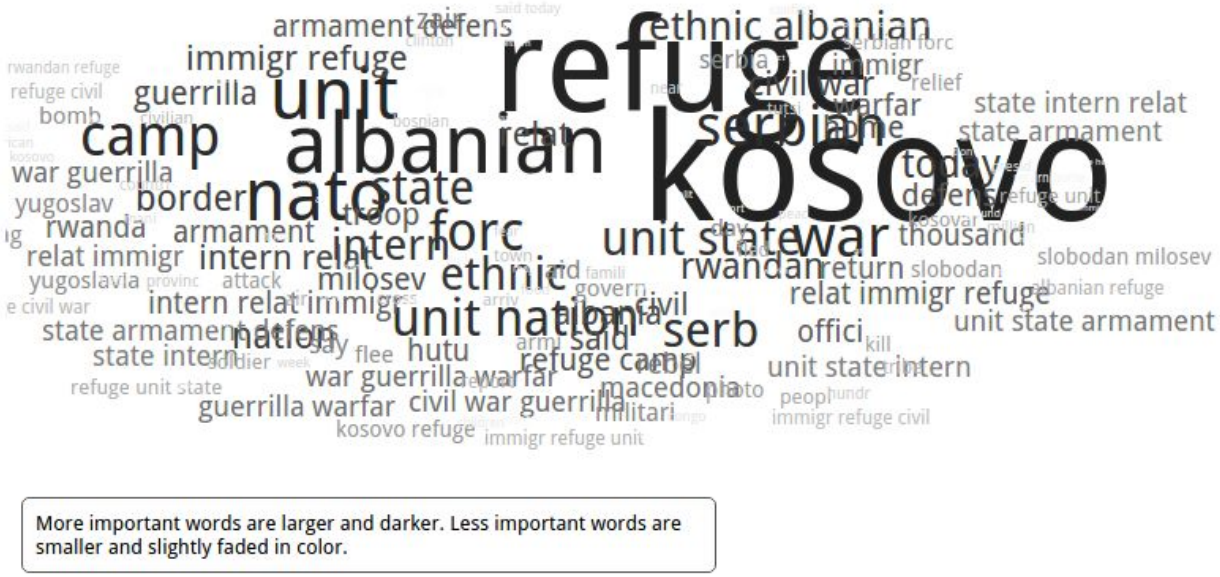


'years', 'states', 'americans', 'new', 'united', 'said', 'government', 'international', 'president',	9,139	58%
'illegal', 'aliens', 'illegal', 'illegal', 'deportation', 'asylum', 'states', 'federal', 'law',	2,843	18%
'cubans', 'cuba', 'elian', 'gonzalez', 'elian', 'relations', 'miami', 'united', 'united', 'states',	710	5%
'york', 'new', 'new', 'city', 'new', 'york', 'schools', 'years', 'book',	1,537	10%
'refugees', 'kosovo', 'united', 'albanian', 'camp', 'war', 'nato', 'forces', 'united', 'serbian',	1,498	10%

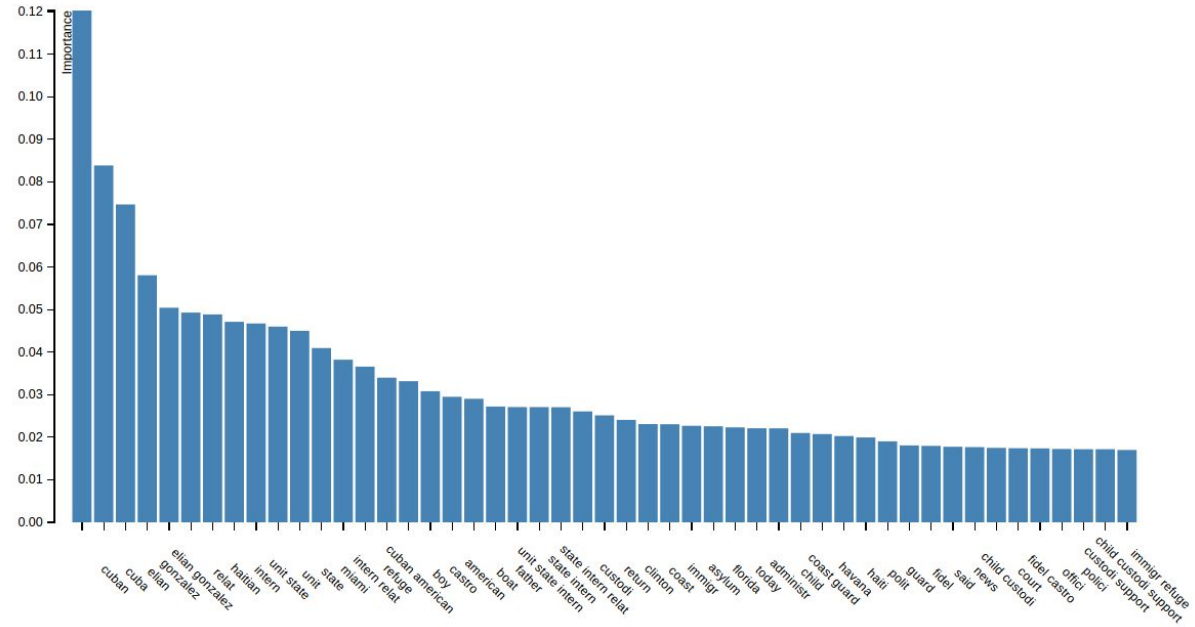
The following is a D3 histogram of the important words in the second cluster by TF-IDF.



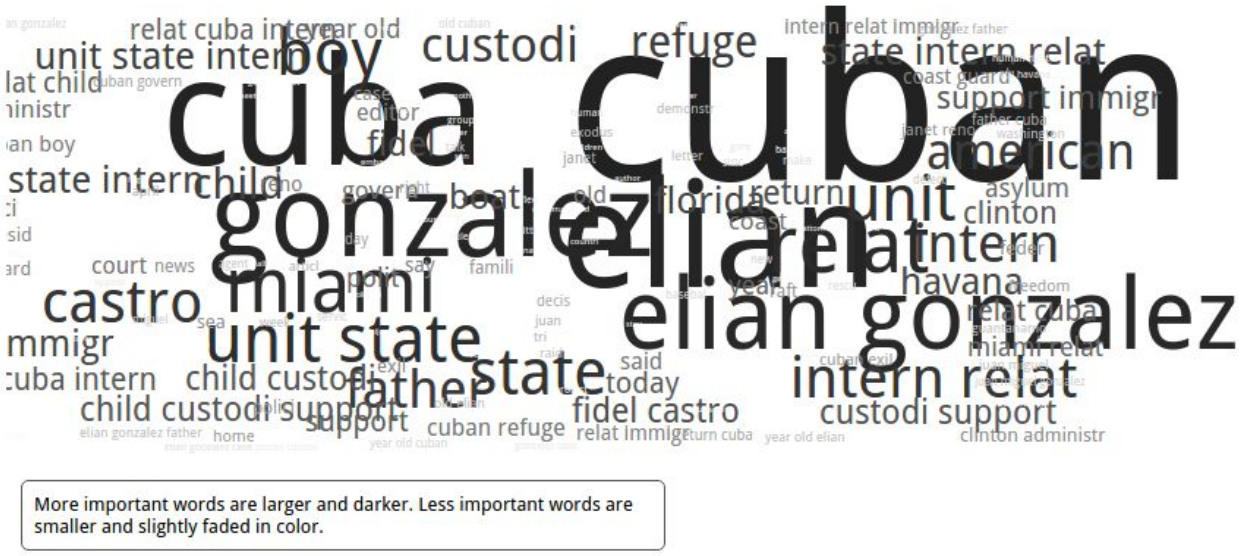
This is a D3 word cloud of the second cluster. As you can see, the word “kosovo” has the highest importance score so its size is bigger and its color is darker. Same as the words like “albanian”, “refugee”, “war” and so on.



The following is a D3 histogram of the important words in the last cluster by TF-IDF



As you can see from the word cloud, the words, “cuban”, “elian gonzalez”, “miami” and “custodi support” show up a lot.



Overall, the clustering tells us two important events, the Kosovo War around 1999 and the Elian Gonzalez custody battle around 2000.