IST565 Data Mining

HW4: Use Clustering to Solve a Mystery in History

In this homework assignment, you are going to use clustering methods to solve a mystery in history: w**ho wrote the disputed essays, Hamilton or Madison?**

1. About the Federalist Papers

Quote from the Library of Congress
http://www.loc.gov/rr/program/bib/ourdocs/federalist.html

The Federalist Papers were a series of eighty-five essays urging the citizens of New York to ratify the new United States Constitution. Written by Alexander Hamilton, James Madison, and John Jay, the essays originally appeared anonymously in New York newspapers in 1787 and 1788 under the pen name "Publius." A bound edition of the essays was first published in 1788, but it was not until the 1818 edition published by the printer Jacob Gideon that the authors of each essay were identified by name. The Federalist Papers are considered one of the most important sources for interpreting and understanding the original intent of the Constitution.

2. About the disputed authorship

The original essays can be downloaded from the Library of Congress.
http://thomas.loc.gov/home/histdox/fedpapers.html

In the author column, you will find 74 essays with identified authors: 51 essays written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by Jay. The remaining 11 essays, however, is authored by "Hamilton or Madison". These are the famous essays with disputed authorship. Hamilton wrote to claim the authorship before he was killed in a duel. Later Madison also claimed authorship. Historians were trying to find out which one was the real author.

3. Computational approach for authorship attribution

In 1960s, statistician Mosteller and Wallace analyzed the frequency distributions of common function words in the Federalist Papers, and drew their conclusions. This is a pioneering work on using mathematical approaches for authorship attribution.
http://www.stat.cmu.edu/~vlachos/courses/724/final/mosteller.pdf

Nowadays, authorship attribution has become a classic problem in the data mining field, with applications in forensics (e.g. deception detection), and information organization.

In this homework you are provided with the Federalist Paper data set. The features are a set of "function words", for example, "upon". The feature value is the percentage of the word occurrence in an essay. For example, for the essay "Hamilton_fed_31.txt", if the function word "upon" appeared 3 times, and the total number of words in this essay is 1000, the feature value is 3/1000=0.3%

Now you are going to try solving this mystery using clustering algorithms k-Means, EM, and HAC. Document your analysis process and draw your conclusion on who wrote the disputed essays. Provide evidence for each method to demonstrate what patterns had been learned to predict the disputed papers, for example, visualize the clustering results and show where the disputed papers are located in relation to Hamilton and Madison's papers. By the way, where are the papers with joint authorship located? For k-Means and EM, analyze the centroids to explain which attributes are most useful for clustering. Hint: the centroid values on these dimensions should be far apart from each other to be able to distinguish the clusters.