Harshit Joshi HW4

IST 707 Applied Machine Learning

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: **who wrote the disputed Federalist 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.
https://guides.loc.gov/federalist-papers/full-text *Note: it is not required that you download these essays. We will work with the data set named **HW4-data-fedPapers85.csv.***

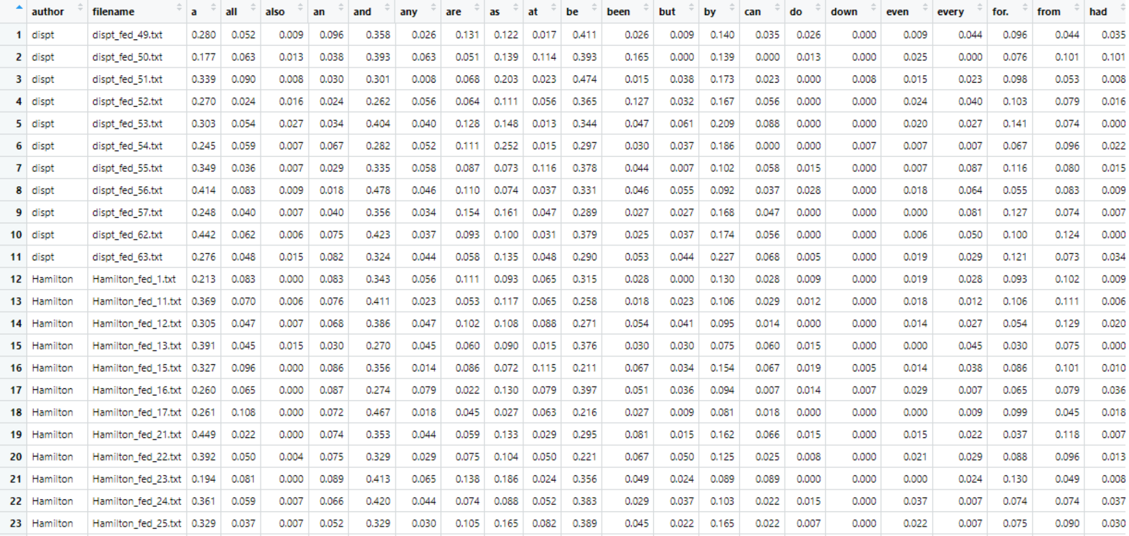
In the author column, you will find 74 essays with identified authors: 51 essays written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by Jay. The remaining 11 essays, however, are authored by “Hamilton or Madison”. These are the famous essays with disputed authorship. Hamilton wrote to claim 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, statisticians 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.

Assignment: 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%**. It is your job to determine the authorship of the disputed essays using**

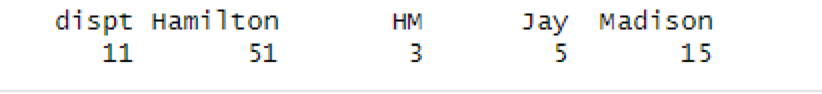
the k-Means clustering algorithm. Document your analysis process and draw your conclusion on who wrote the disputed essays. Provide evidence for each method to demonstrate what patterns had been learned to predict the disputed papers, for example, visualize the clustering results and show where the disputed papers are located in relation to Hamilton and Madison's papers. 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.



	author	filename	a	all	also	an	and	any	are	as	at	be	been	but	by	can	do	down	even	every	for.	from	had
1	dispt	dispt_fed_49.txt	0.280	0.052	0.009	0.096	0.358	0.026	0.131	0.122	0.017	0.411	0.026	0.009	0.140	0.035	0.026	0.000	0.009	0.044	0.096	0.044	0.035
2	dispt	dispt_fed_50.txt	0.177	0.063	0.013	0.038	0.393	0.063	0.051	0.139	0.114	0.393	0.165	0.000	0.139	0.000	0.013	0.000	0.025	0.000	0.076	0.101	0.101
3	dispt	dispt_fed_51.txt	0.339	0.090	0.008	0.030	0.301	0.008	0.068	0.203	0.023	0.474	0.015	0.038	0.173	0.023	0.000	0.008	0.015	0.023	0.098	0.053	0.008
4	dispt	dispt_fed_52.txt	0.270	0.024	0.016	0.024	0.262	0.056	0.064	0.111	0.056	0.365	0.127	0.032	0.167	0.056	0.000	0.000	0.024	0.040	0.103	0.079	0.016
5	dispt	dispt_fed_53.txt	0.303	0.054	0.027	0.034	0.404	0.040	0.128	0.148	0.013	0.344	0.047	0.061	0.209	0.088	0.000	0.000	0.020	0.027	0.141	0.074	0.000
6	dispt	dispt_fed_54.txt	0.245	0.059	0.007	0.067	0.282	0.052	0.111	0.252	0.015	0.297	0.030	0.037	0.186	0.000	0.000	0.007	0.007	0.007	0.067	0.096	0.022
7	dispt	dispt_fed_55.txt	0.349	0.036	0.007	0.029	0.335	0.058	0.087	0.073	0.116	0.378	0.044	0.007	0.102	0.058	0.015	0.000	0.007	0.087	0.116	0.080	0.015
8	dispt	dispt_fed_56.txt	0.414	0.083	0.009	0.018	0.478	0.046	0.110	0.074	0.037	0.331	0.046	0.055	0.092	0.037	0.028	0.000	0.018	0.064	0.055	0.083	0.009
9	dispt	dispt_fed_57.txt	0.248	0.040	0.007	0.040	0.356	0.034	0.154	0.161	0.047	0.289	0.027	0.027	0.168	0.047	0.000	0.000	0.000	0.081	0.127	0.074	0.007
10	dispt	dispt_fed_62.txt	0.442	0.062	0.006	0.075	0.423	0.037	0.093	0.100	0.031	0.379	0.025	0.037	0.174	0.056	0.000	0.000	0.006	0.050	0.100	0.124	0.000
11	dispt	dispt_fed_63.txt	0.276	0.048	0.015	0.082	0.324	0.044	0.058	0.135	0.048	0.290	0.053	0.044	0.227	0.068	0.005	0.000	0.019	0.029	0.121	0.073	0.034
12	Hamilton	Hamilton_fed_1.txt	0.213	0.083	0.000	0.083	0.343	0.056	0.111	0.093	0.065	0.315	0.028	0.000	0.130	0.028	0.009	0.000	0.019	0.028	0.093	0.102	0.009
13	Hamilton	Hamilton_fed_11.txt	0.369	0.070	0.006	0.076	0.411	0.023	0.053	0.117	0.065	0.258	0.018	0.023	0.106	0.029	0.012	0.000	0.018	0.012	0.106	0.111	0.006
14	Hamilton	Hamilton_fed_12.txt	0.305	0.047	0.007	0.068	0.386	0.047	0.102	0.108	0.088	0.271	0.054	0.041	0.095	0.014	0.000	0.000	0.014	0.027	0.054	0.129	0.020
15	Hamilton	Hamilton_fed_13.txt	0.391	0.045	0.015	0.030	0.270	0.045	0.060	0.090	0.015	0.376	0.030	0.030	0.075	0.060	0.015	0.000	0.000	0.045	0.030	0.075	0.000
16	Hamilton	Hamilton_fed_15.txt	0.327	0.096	0.000	0.086	0.356	0.014	0.086	0.072	0.115	0.211	0.067	0.034	0.154	0.067	0.019	0.005	0.014	0.038	0.086	0.101	0.010
17	Hamilton	Hamilton_fed_16.txt	0.260	0.065	0.000	0.087	0.274	0.079	0.022	0.130	0.079	0.397	0.051	0.036	0.094	0.007	0.014	0.007	0.029	0.007	0.065	0.079	0.036
18	Hamilton	Hamilton_fed_17.txt	0.261	0.108	0.000	0.072	0.467	0.018	0.045	0.027	0.063	0.216	0.027	0.009	0.081	0.018	0.000	0.000	0.000	0.009	0.099	0.045	0.018
19	Hamilton	Hamilton_fed_21.txt	0.449	0.022	0.000	0.074	0.353	0.044	0.059	0.133	0.029	0.295	0.081	0.015	0.162	0.066	0.015	0.000	0.015	0.022	0.037	0.118	0.007
20	Hamilton	Hamilton_fed_22.txt	0.392	0.050	0.004	0.075	0.329	0.029	0.075	0.104	0.050	0.221	0.067	0.050	0.125	0.025	0.008	0.000	0.021	0.029	0.088	0.096	0.013
21	Hamilton	Hamilton_fed_23.txt	0.194	0.081	0.000	0.089	0.413	0.065	0.138	0.186	0.024	0.356	0.049	0.024	0.089	0.089	0.000	0.000	0.000	0.024	0.130	0.049	0.008
22	Hamilton	Hamilton_fed_24.txt	0.361	0.059	0.007	0.066	0.420	0.044	0.074	0.088	0.052	0.383	0.029	0.037	0.103	0.022	0.015	0.000	0.037	0.007	0.074	0.074	0.037
23	Hamilton	Hamilton_fed_25.txt	0.329	0.037	0.007	0.052	0.329	0.030	0.105	0.165	0.082	0.389	0.045	0.022	0.165	0.022	0.007	0.000	0.022	0.007	0.075	0.090	0.030

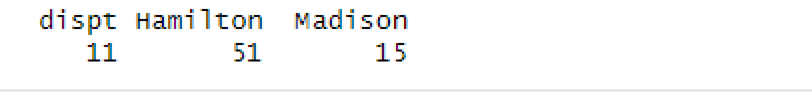
Steps:

- **Load the csv file as a data frame Essay_main. Once loaded I found that our dataset contains information about 85 authors (72 words) in which 11 were dispt values meaning disputed authors. Other authors were Hamilton, Madison, Jay and HM ~ Hamilton and Madison.**
- **In order to find out which author wrote the disputed essay, I need to run k-means clustering on the data set and analyze the two clusters. Here I don't need author – Jay and HM as they are not important and might be outliers in kmeans. Also the number of clusters I choose should be 2.**



```
   dispt Hamilton       HM      Jay  Madison 
      11       51        3        5       15 
```

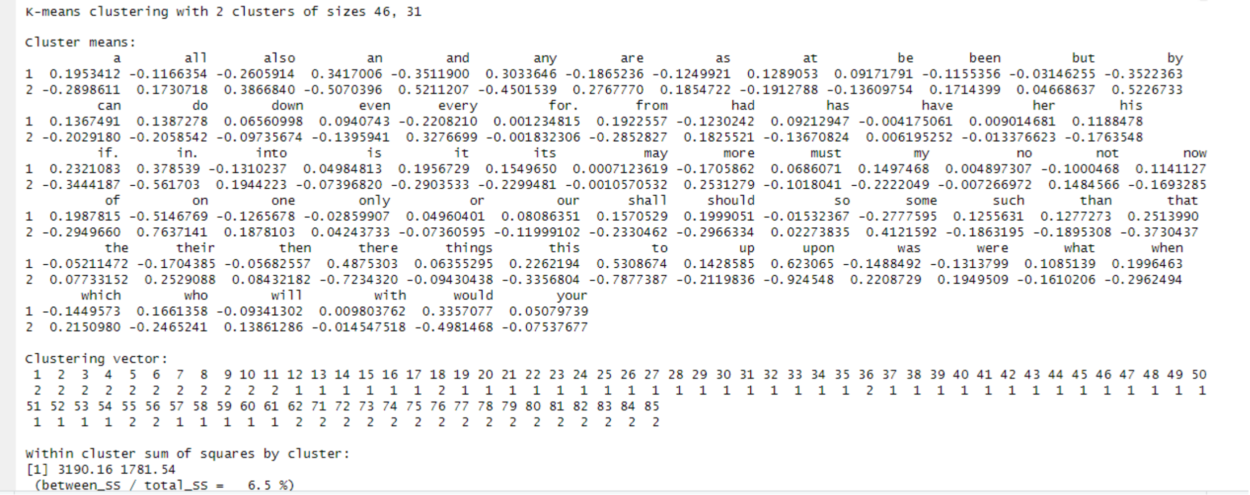
After Removing HM and Jay:



```
  dispt Hamilton  Madison 
     11       51       15 
```

- **To run kmeans, I need data in form of matrix, therefore I need to convert my data to matrix. But for this I will need to remove catagorical values from my dataset. Therefore I removed the author and filename column and stored it in new variable called Essay_main.unlabeled**

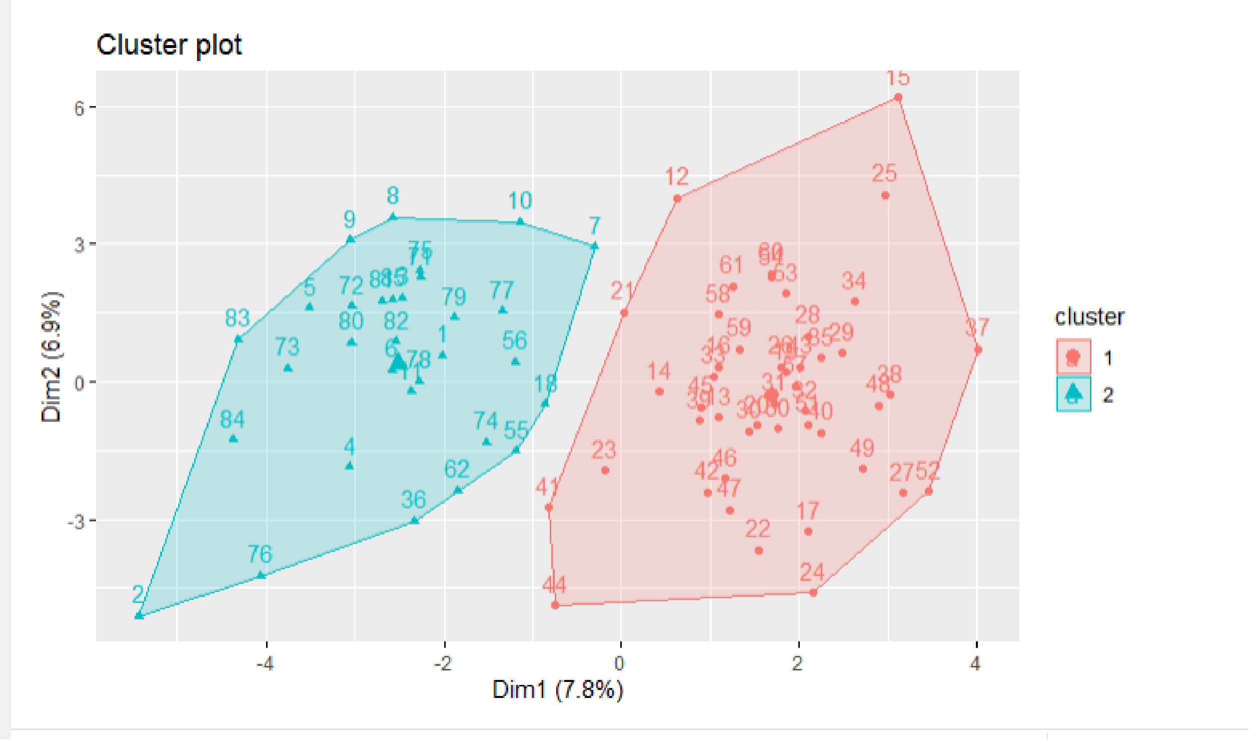
- **After this I’m ready to scale my data using scale() function. It will convert my unlabeled data to matrix form.**
- **Now I put my unlabeled data in kmeans() and analyze the values for each cluster, each word**

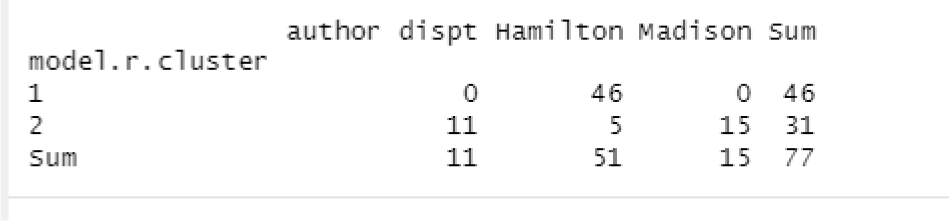


```
K-means clustering with 2 clusters of sizes 46, 31

Cluster means:
           a        all       also         an        and        any        are         as         at          be       been         but         by
1  0.1953412 -0.1166354 -0.2605914  0.3417006 -0.3511900  0.3033646 -0.1865236 -0.1249921  0.1289053  0.09171791 -0.1155356 -0.03146255 -0.3522363
2 -0.2898611  0.1730718  0.3866840 -0.5070396  0.5211207 -0.4501539  0.2767770  0.1854722 -0.1912788 -0.13609754  0.1714399  0.04668637  0.5226733
         can         do        down       even      every         for.       from         had         has         have          her        his
1  0.1367491  0.1387278  0.06560998  0.0940743 -0.2208210  0.001234815  0.1922557 -0.1230242  0.09212947 -0.004175061  0.009014681  0.1188478
2 -0.2029180 -0.2058542 -0.09735674 -0.1395941  0.3276699 -0.001832306 -0.2852827  0.1825521 -0.13670824  0.006195252 -0.013376623 -0.1763548
         if.        in.       into          is          it        its           may       more       must          my           no        not        now
1  0.2321083  0.378539 -0.1310237  0.04984813  0.1956729  0.1549650  0.0007123619 -0.1705862  0.0686071  0.1497468  0.004897307 -0.1000468  0.1141127
2 -0.3444187 -0.561703  0.1944223 -0.07396820 -0.2903533 -0.2299481 -0.0010570532  0.2531279 -0.1018041 -0.2222049 -0.007266972  0.1484566 -0.1693285
          of         on        one        only          or         our      shall     should          so       some       such      than       that
1  0.1987815 -0.5146769 -0.1265678 -0.02859907  0.04960401  0.08086351  0.1570529  0.1999051 -0.01532367 -0.2777595  0.1255631  0.1277273  0.2513990
2 -0.2949660  0.7637141  0.1878103  0.04243733 -0.07360595 -0.11999102 -0.2330462 -0.2966334  0.02273835  0.4121592 -0.1863195 -0.1895308 -0.3730437
          the       their        then       there      things       this         to         up       upon        was       were       what       when
1 -0.05211472 -0.1704385 -0.05682557  0.4875303  0.06355295  0.2262194  0.5308674  0.1428585  0.623065 -0.1488492 -0.1313799  0.1085139  0.1996463
2  0.07733152  0.2529088  0.08432182 -0.7234320 -0.09430438 -0.3356804 -0.7877387 -0.2119836 -0.924548  0.2208729  0.1949509 -0.1610206 -0.2962494
       which        who        will         with      would       your
1 -0.1449573  0.1661358 -0.09341302  0.009803762  0.3357077  0.05079739
2  0.2150980 -0.2465241  0.13861286 -0.014547518 -0.4981468 -0.07537677

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 
 2  2  2  2  2  2  2  2  2  2  2  1  1  1  1  1  1  2  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  1  1  1  1  1  1  1  1  1  1  1  1  1  1 
51 52 53 54 55 56 57 58 59 60 61 62 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 
 1  1  1  1  2  2  1  1  1  1  1  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2 

Within cluster sum of squares by cluster:
[1] 3190.16 1781.54
 (between_SS / total_SS =   6.5 %)
```

- **As the clusters are formed, I can see that values of each word are on the other side of zero for both clusters. This shows that each word is useful in formation of clusters. If I see further, words like (there, upon, to, on) have highest difference between the 2 clusters compare to other words. It can be said that these words are more helpful for us in clustering.**

- **This is how the two clusters look, one of these clusters represents Hamilton and the other Madison. And we must have dispt values situated within one of these clusters:**





```
                 author dispt Hamilton Madison Sum
model.r.cluster
1                           0       46       0  46
2                          11        5      15  31
Sum                        11       51      15  77
```

- **By viewing at this table, it looks like cluster 2 (lower row) contains all the dispt values, some Hamilton values and almost all of Madison values. Whereas Cluster 1 (upper row) only contains Hamilton values. Therefore it can be said that the writer of the essay was Madison as the cluster with Madison contains all dispt values.**