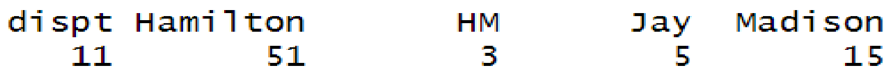
IST707 Applied Machine Learning
School of Information Studies
Syracuse University

Hendi Kushta

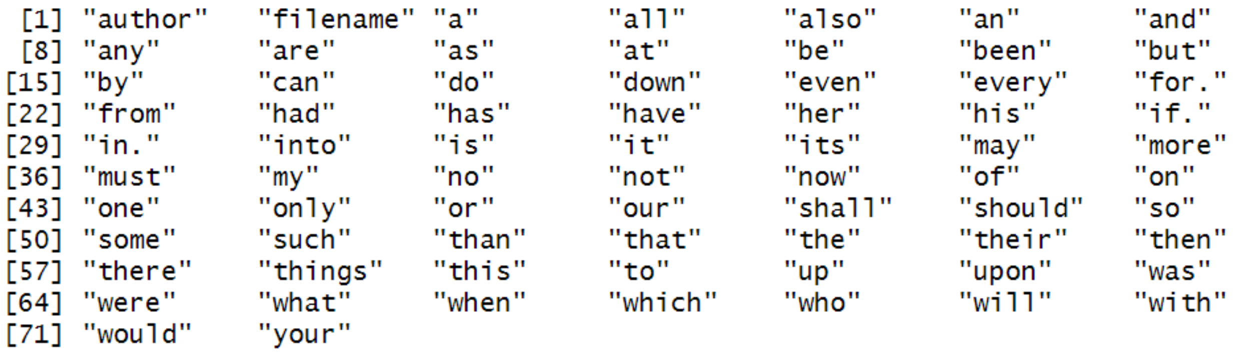
HW4: ClusteringQ 1: Who wrote the disputed essays, Hamilton or Madison?
In order to identify the author of the essays we can analyze the characteristics and the writing styles of the authors. This analysis starts by verifying the given dataset, some exploratory analysis is also performed, and finally we analyze through unsupervised learning (with k-means algorithm and hierarchical) the possibility of authorship for the essays.

R code written in RStudio is used to develop the project. When the dataset is imported from the csv file into the project, 85 rows (observations) and 72 columns are seen (attributes). The columns are written essays, while the qualities are made up of "function words." The distribution of these essays among their writers can be seen using the table function:



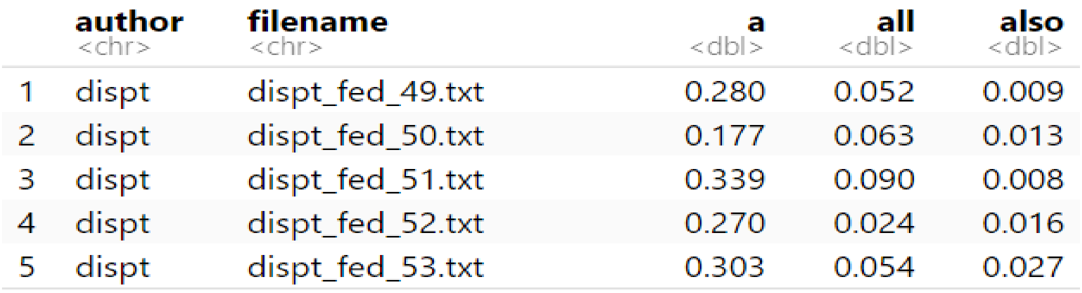
```
   dispt Hamilton       HM      Jay  Madison 
      11       51        3        5       15 
```

51 writings by Hamilton, 5 by Jay, 15 by Madison, and 3 by Hamilton and Madison are included in this dataset. The 11 essays that are left are the ones whose authorship is under question. The following is a list of all function words:



```
 [1] "author"   "filename" "a"        "all"      "also"     "an"       "and"     
 [8] "any"      "are"      "as"       "at"       "be"       "been"     "but"     
[15] "by"       "can"      "do"       "down"     "even"     "every"    "for."    
[22] "from"     "had"      "has"      "have"     "her"      "his"      "if."     
[29] "in."      "into"     "is"       "it"       "its"      "may"      "more"    
[36] "must"     "my"       "no"       "not"      "now"      "of"       "on"      
[43] "one"      "only"     "or"       "our"      "shall"    "should"   "so"      
[50] "some"     "such"     "than"     "that"     "the"      "their"    "then"    
[57] "there"    "things"   "this"     "to"       "up"       "upon"     "was"     
[64] "were"     "what"     "when"     "which"    "who"      "will"     "with"    
[71] "would"    "your"    
```

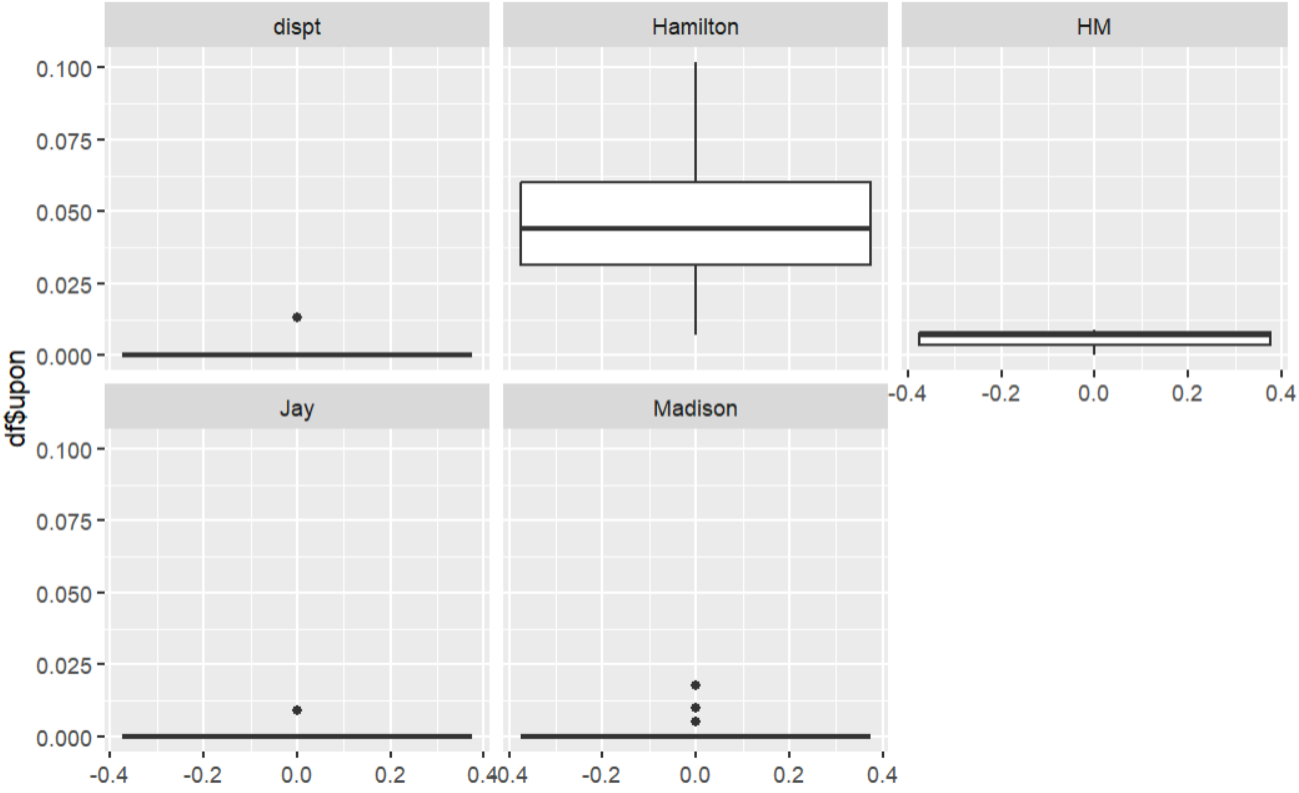
In order to avoid giving some of the attributes a higher weight than the others, the author and filename columns must be removed, and data normalization must be performed. As a result, all of the attributes should have a mean of 0 and a standard deviation of 1. An illustration of normalizing before and after is shown below:



	author <chr>	filename <chr>	a <dbl>	all <dbl>	also <dbl>
1	dispt	dispt_fed_49.txt	0.280	0.052	0.009
2	dispt	dispt_fed_50.txt	0.177	0.063	0.013
3	dispt	dispt_fed_51.txt	0.339	0.090	0.008
4	dispt	dispt_fed_52.txt	0.270	0.024	0.016
5	dispt	dispt_fed_53.txt	0.303	0.054	0.027

My next step was to check if there is any missing value or different from numerical and there is not any.

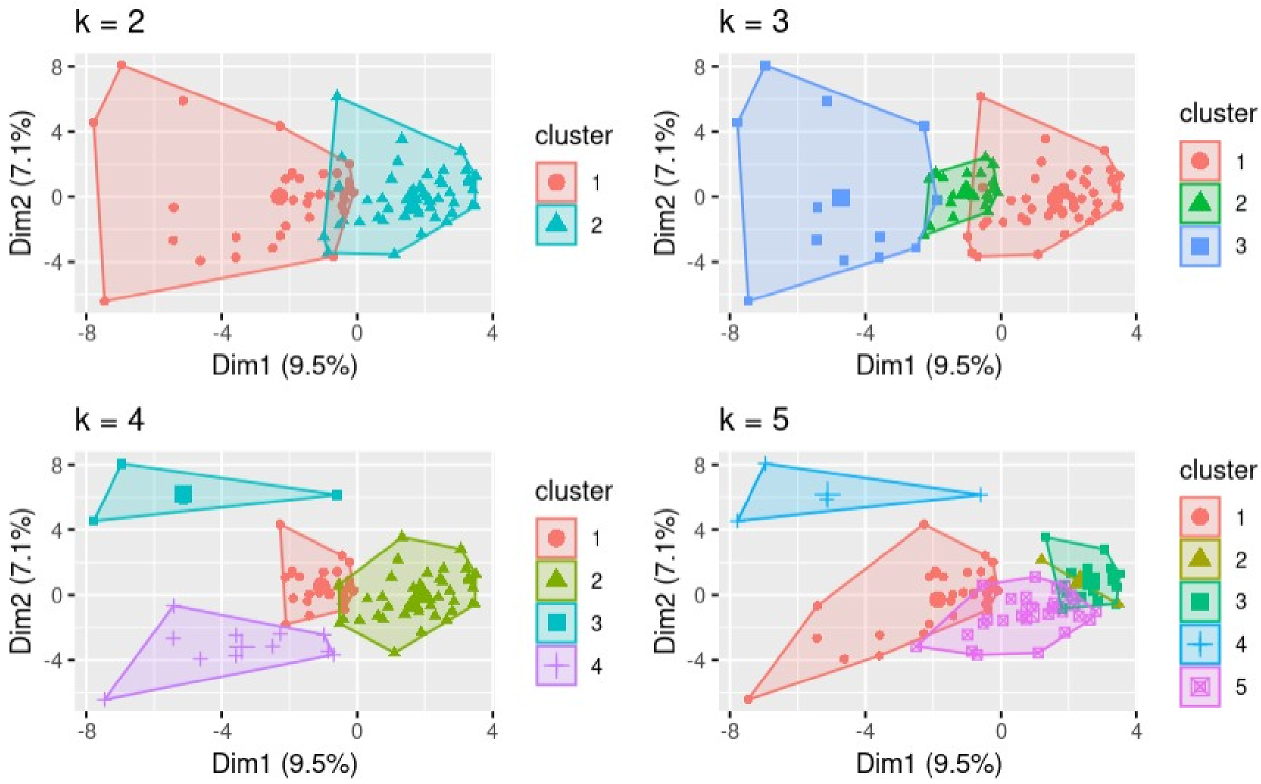
Simple exploratory analysis allows us to determine how different authors differ from one another. We may see the differences, for instance, by selecting a sample of three function words and displaying their distribution using a boxplot. The phrases "all onto" and "into" have been chosen. The boxplot of "upon" words demonstrates that Hamilton uses this word far more frequently than the others as a writer:



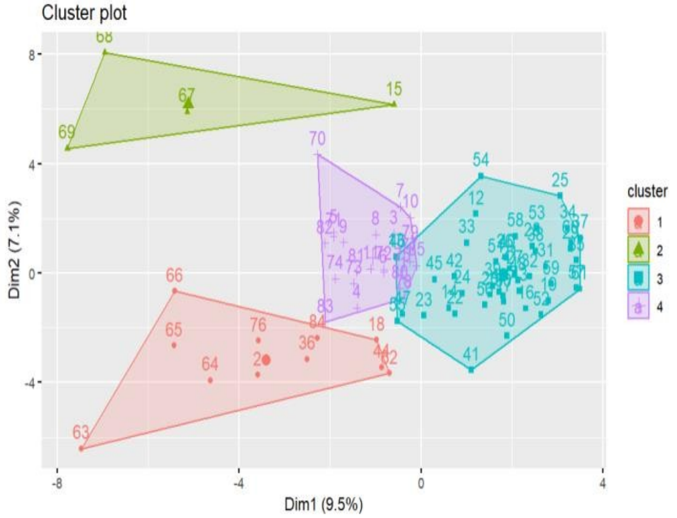
Our dataset contains 72 functional words; thus, we would need a clustering algorithm to identify and classify similar patterns.

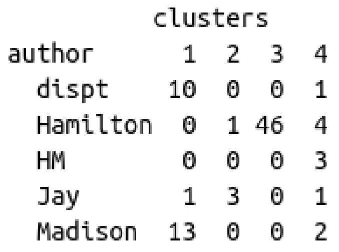
K-Means

K-means, which conducts a hard categorization of the observation, is one of the most popular unsupervised classification methods. This implies that each observation is a member of a distinct cluster. Finding the right number of clusters is one of k-means' challenges; as a result, a number of variables, including domain knowledge, can be taken into account. The number of times the algorithm is run with different initial starting points for the cluster centroids is 25 and 4 centroids where taken in consideration.
After performing some visualizations, we can see the different clustering graphs and see that for k=2,3 and 5 the clusters have several overlapping compared to k=4. Thus, it is decided to proceed with the analysis of finding the authors and clustering in 4 categories.

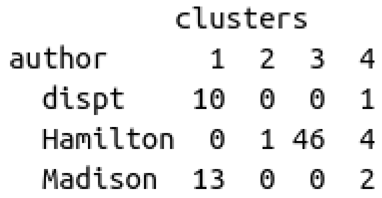


Performing the k-means modeling, we can identify the four clusters but yet we have to check on which cluster are categorized our disputed papers. For this, it is necessary to map the cluster number with the author names.





```
              clusters
author        1  2  3  4
  dispt      10  0  0  1
  Hamilton    0  1 46  4
  HM          0  0  0  3
  Jay         1  3  0  1
  Madison    13  0  0  2
```



```
           clusters
author      1  2  3  4
  dispt    10  0  0  1
  Hamilton  0  1 46  4
  Madison  13  0  0  2
```

Since the question is weather the essays are written from Hamilton or Madison, from the above clusters, there is a high chance that Madison has written them, since 10 of the disputed essays are in the first cluster which is Madison’s Cluster.