Jacob Dineen

IST 565 Data Mining

Homework 4

10/31/17

**Using Weka**

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

74 total essay have authorship identified by historians, as noted below by segment.

Hamilton: 51

Madison: 15

Hamilton and Madison: 3

Jay: 5

11 essays are marked as disputed, meaning that the original authorship is subject to interpretation because both Hamilton and Madison claimed the works.

We are going to use word usage as a measurement of authorship amongst the instances provided. Our hope is to see a correlation amongst the feature value of a particular word, defined as the total number of appearances of a word divided by the total word count of the essay, and the author of the essay. We will withhold/ignore the attributes ‘Author’ and ‘Filename’ from our clustering, noting that the Author is withheld from the computation when using the classes to cluster function within Weka.

We will use a seed of 10 throughout this analysis process for the K Means Clusterer, as the number of incorrectly clustered instances fluctuates drastically when moving in either direction.

Also, we will use a K=2, meaning we will be assigning all instances to two different clusters. When K >3, we have more incorrectly assigned instances to class type. Class type is also identified as Author.

**Simple K Means**

Starting with the Simple K Means algorithm, utilizing Euclidean Distance, we see the following:

kMeans

Number of iterations: 4

Within cluster sum of squared errors: 235.4860832140816

=== Model and evaluation on training set ===

Clustered Instances

0 37 ( 44%)

1 48 ( 56%)

Class attribute: author

Classes to Clusters:

0 1 <-- assigned to cluster

11 0 | dispt

3 48 | Hamilton

3 0 | HM

5 0 | Jay

15 0 | Madison

Cluster 0 <-- Madison

Cluster 1 <-- Hamilton

Incorrectly clustered instances : 22.0 25.8824 %

From the Simple K Means method, we can see that all 11 of the disputed essays use language most related to the workings of Madison. We can also see that:

1. Jay used words most related to Madison, opposed to Hamilton.

2. The joint essays were most likely written, atleast in large part, by Madison – All of the HM class was thrown into cluster 0, or Class ‘Madison’.

3. Three of Hamilton’s essays were heavy in diction/language most used by Madison.

**Heirarchical Clusterer**

Next, we will use the Heirarchical Clusterer. Using K=2 here was not helpful, as the clusters were defined as Class = Hamilton, and Class = Jay. Increasing our number of clusters resulted in the following:

=== Model and evaluation on training set ===

Clustered Instances

0 75 ( 88%)

1 1 ( 1%)

2 6 ( 7%)

3 3 ( 4%)

Class attribute: author

Classes to Clusters:

0 1 2 3 <-- assigned to cluster

11 0 0 0 | dispt

50 1 0 0 | Hamilton

0 0 3 0 | HM

0 0 2 3 | Jay

14 0 1 0 | Madison

Cluster 0 <-- Hamilton

Cluster 1 <-- No class

Cluster 2 <-- HM

Cluster 3 <-- Jay

Incorrectly clustered instances : 29.0 34.1176 %

We can see through the confusion matrix that Madison, Hamilton and Disputed were all lumped into cluster 1. This Clusterer was unable to identify the difference in linguistic stylings between the relevant authors. We will not rely heavily on Heirarchical Clustering for this assignment.

**EM**

Using the EM Clusterer arguably produced our most fulfilling results.

=== Model and evaluation on training set ===

Clustered Instances

0 34 ( 40%)

1 51 ( 60%)

Log likelihood: 161.83817

Class attribute: author

Classes to Clusters:

0 1 <-- assigned to cluster

11 0 | dispt

0 51 | Hamilton

3 0 | HM

5 0 | Jay

15 0 | Madison

Cluster 0 <-- Madison

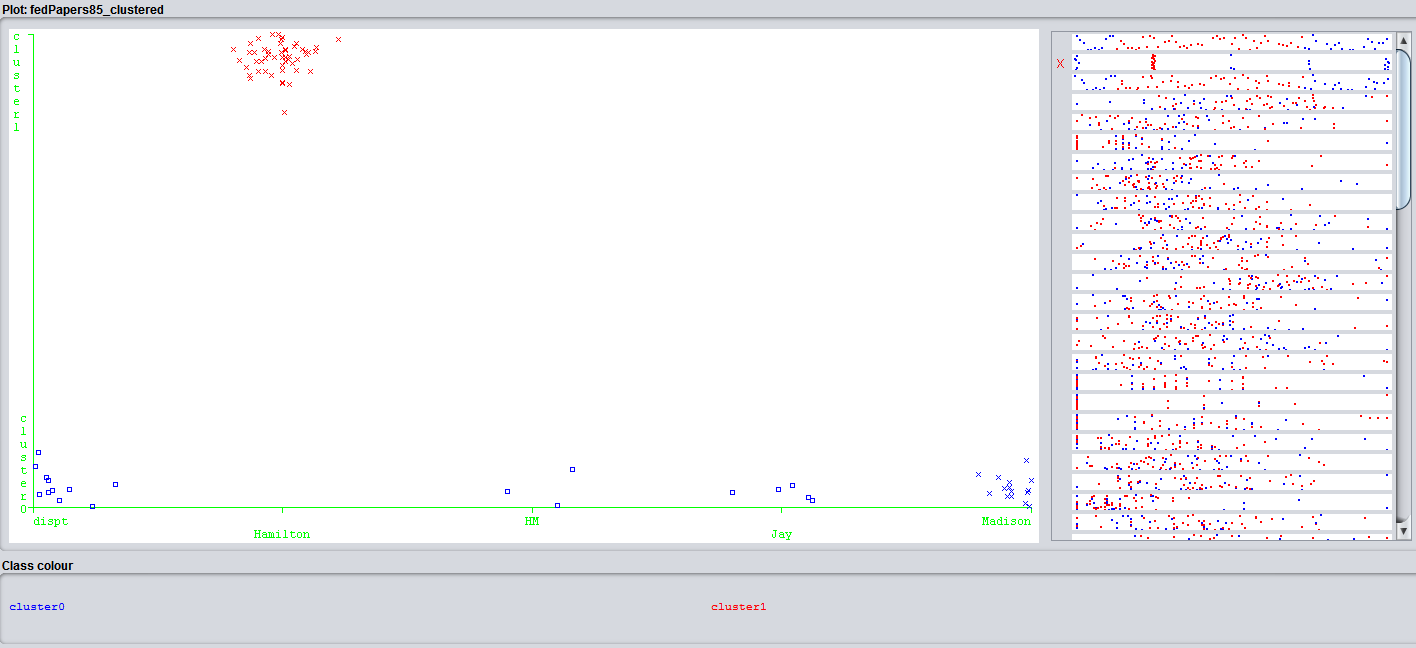
Cluster 1 <-- Hamilton

Incorrectly clustered instances : 19.0 22.3529 %

Very much along the lines of the Simple K Means Clusterer, except this model was able to successful identify each of Hamilton’s essays. The points from above still apply:

1. Jay used words most related to Madison, opposed to Hamilton.

2. The joint essays were most likely written, at least in large part, by Madison – All of the HM class was thrown into cluster 0, or Class ‘Madison’.



As seen above, 51 instances appear in cluster two, noted in red. The remaining instances fall into the cluster 1 (True 0) bucket. This includes all three of the joint papers.

Because our attributes are our word values, we can look at the centroids between the two clusters, as associated with each attribute.

When looking at the Simple K Means centroids, we consider the algorithm that gets us there:

sqrt(sum((a-b)^2))

Initial starting points (random):

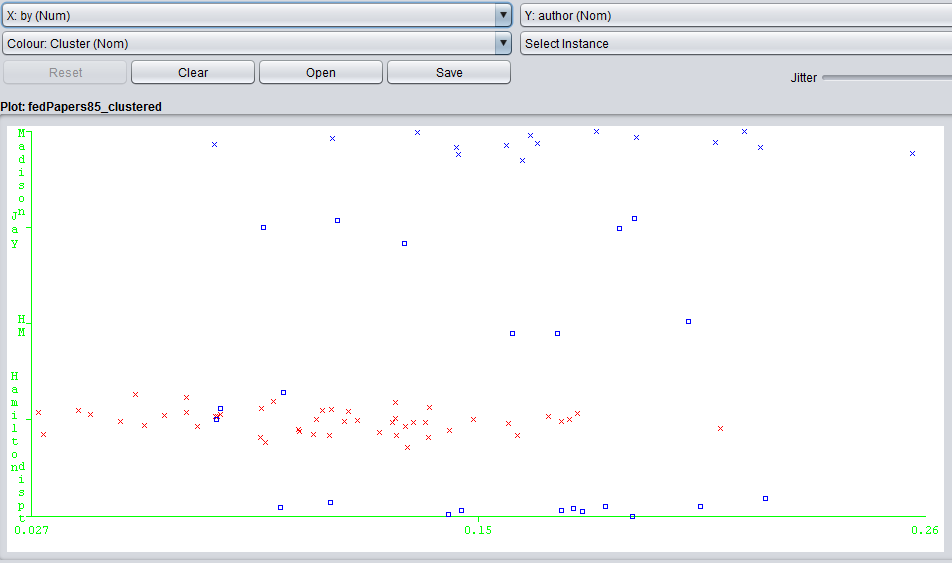
Cluster 0: 0.27,0.024,0.016,0.024,0.262,0.056,0.064,0.111,0.056,0.365,0.127,0.032,0.167,0.056,0,0,0.024,0.04,0.103,0.079,0.016,0.024,0.143,0,0.024,0.04,0.238,0.008,0.151,0.222,0.048,0.056,0.056,0.071,0,0.032,0.087,0,0.802,0.143,0.032,0.048,0.064,0.016,0.016,0.032,0.04,0.024,0.008,0,0.238,1.326,0.071,0.008,0,0,0.103,0.532,0,0,0.087,0.079,0.008,0.024,0.167,0,0.079,0.079,0.064,0

Cluster 1: 0.349,0.066,0,0.079,0.217,0.079,0.066,0.125,0.053,0.441,0.013,0.02,0.138,0.026,0,0,0,0.013,0.079,0.079,0,0.02,0.04,0,0,0.02,0.494,0.033,0.178,0.25,0.066,0.066,0.026,0.04,0,0.026,0.086,0.007,0.962,0.033,0.04,0.007,0.105,0,0.013,0.033,0.04,0.02,0.026,0.046,0.303,1.416,0.092,0.007,0.046,0.007,0.112,0.573,0,0.053,0,0.013,0.033,0,0.105,0.046,0.132,0.04,0.165,0

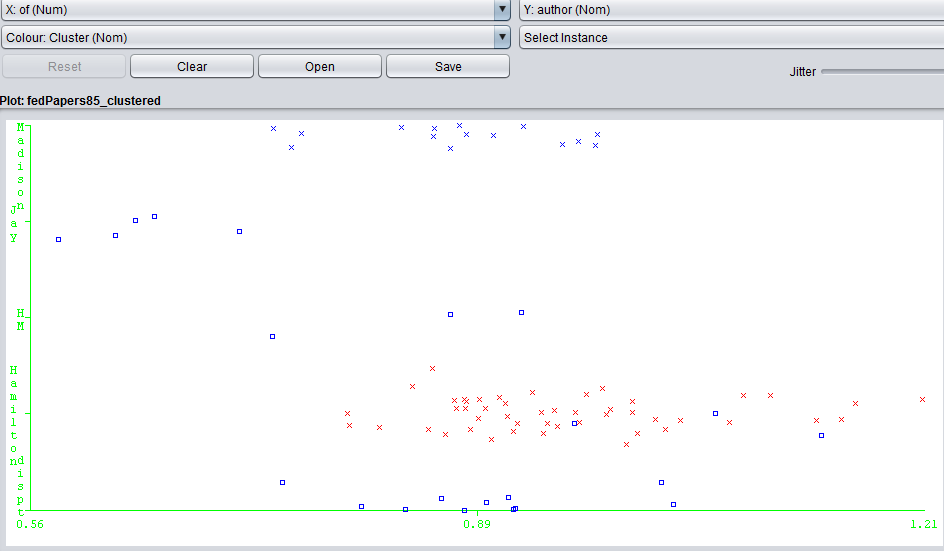
When looking at the visuals for the Simple K Means Clusterer, we can see that there are immensely more instances of attributes for cluster 1, as opposed to cluster 0, mainly due to the number of files that are associated with each cluster. This would seem to signify that the centroids of cluster 1 are less impacted by outliers – As the range of a set of values will not drastically impact the feature value associated with an attribute.

Visualizing high value cluster centroids (initially using 0.10 as a baseline), we can start to see some patterns.

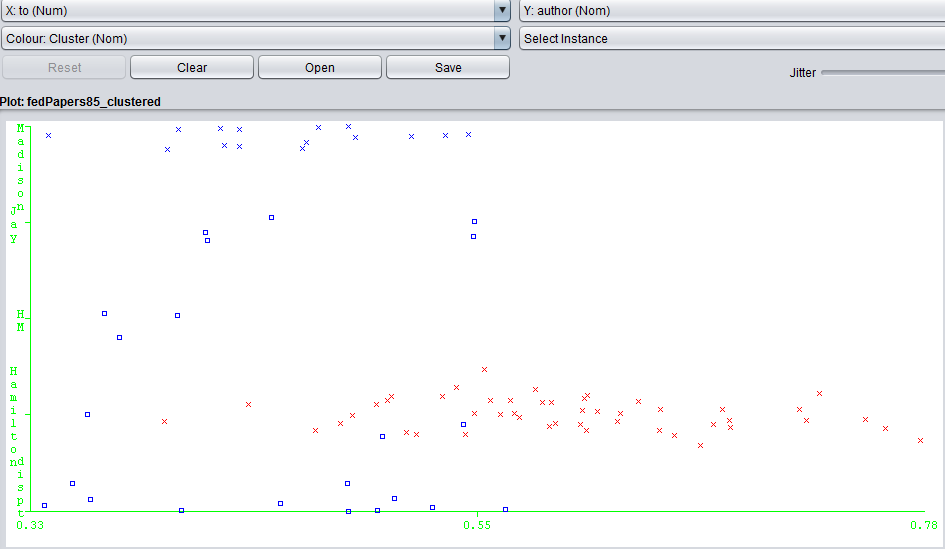
The word by occurs much more frequently in cluster one then it does in cluster two. The Disputed Author column shows us that the feature values of this attribute is more inline with the centroids of Madison’s workings.

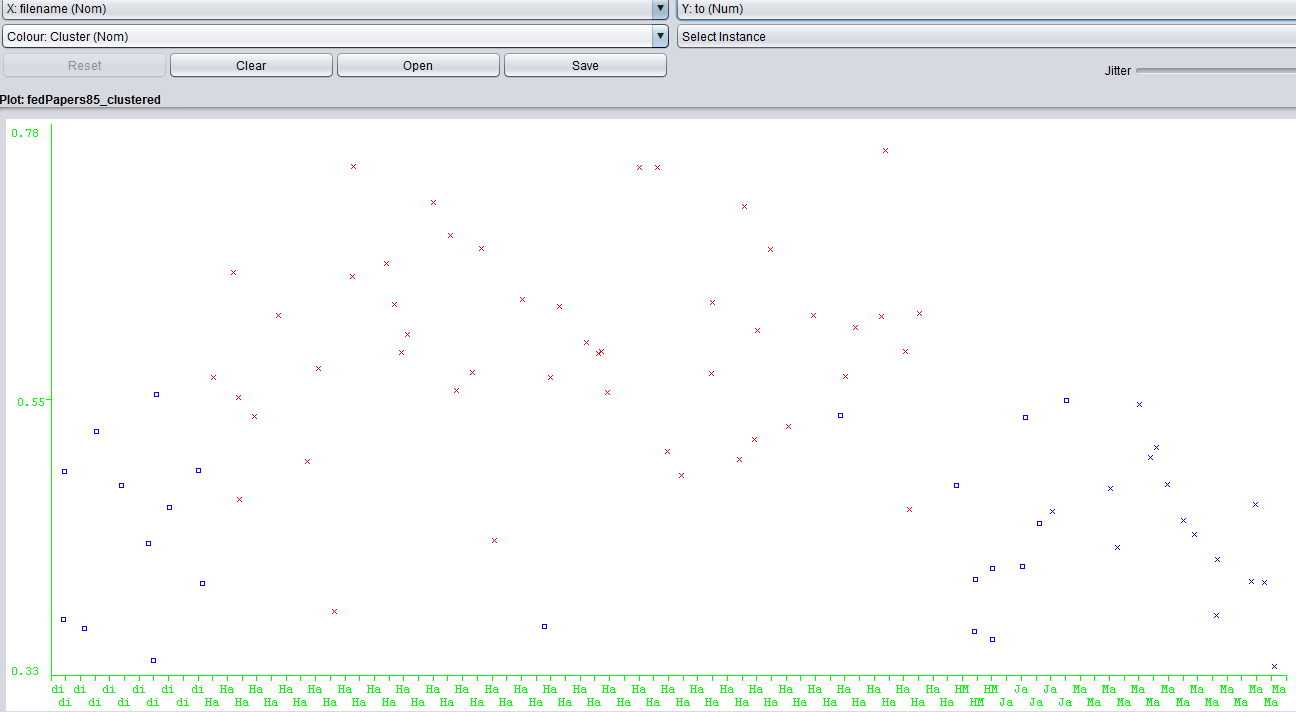


Similar can be said of the function word ‘of’ :



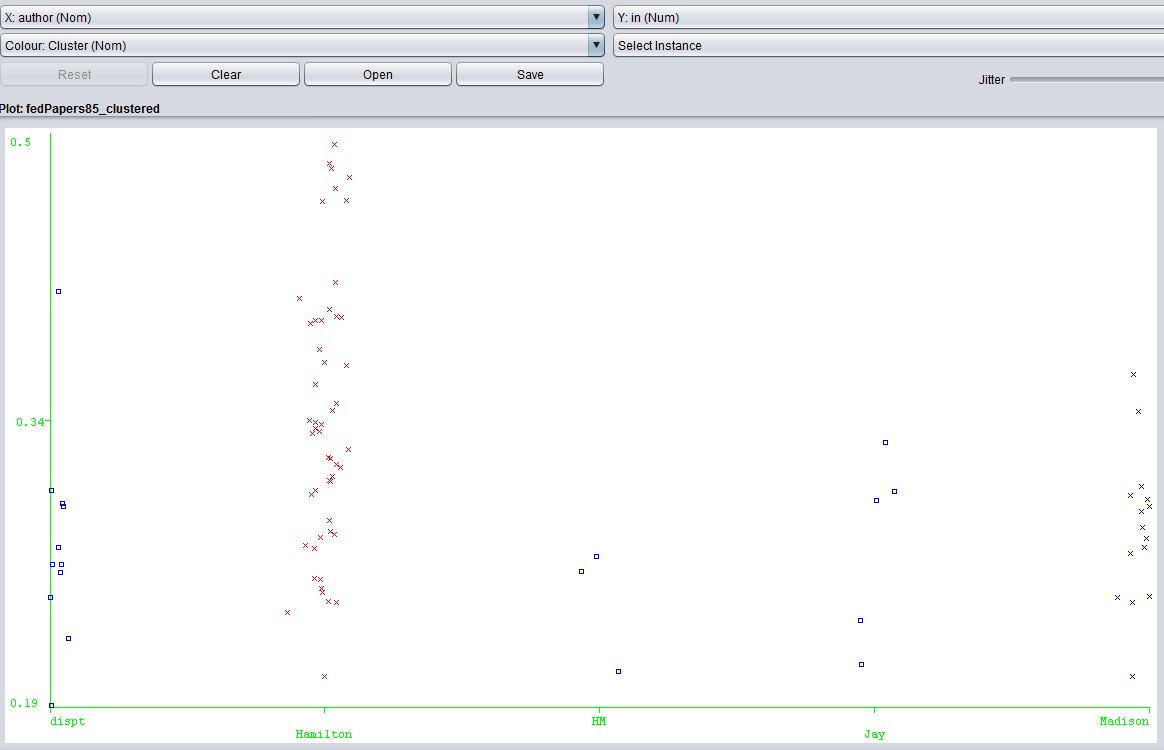
And ‘to’:



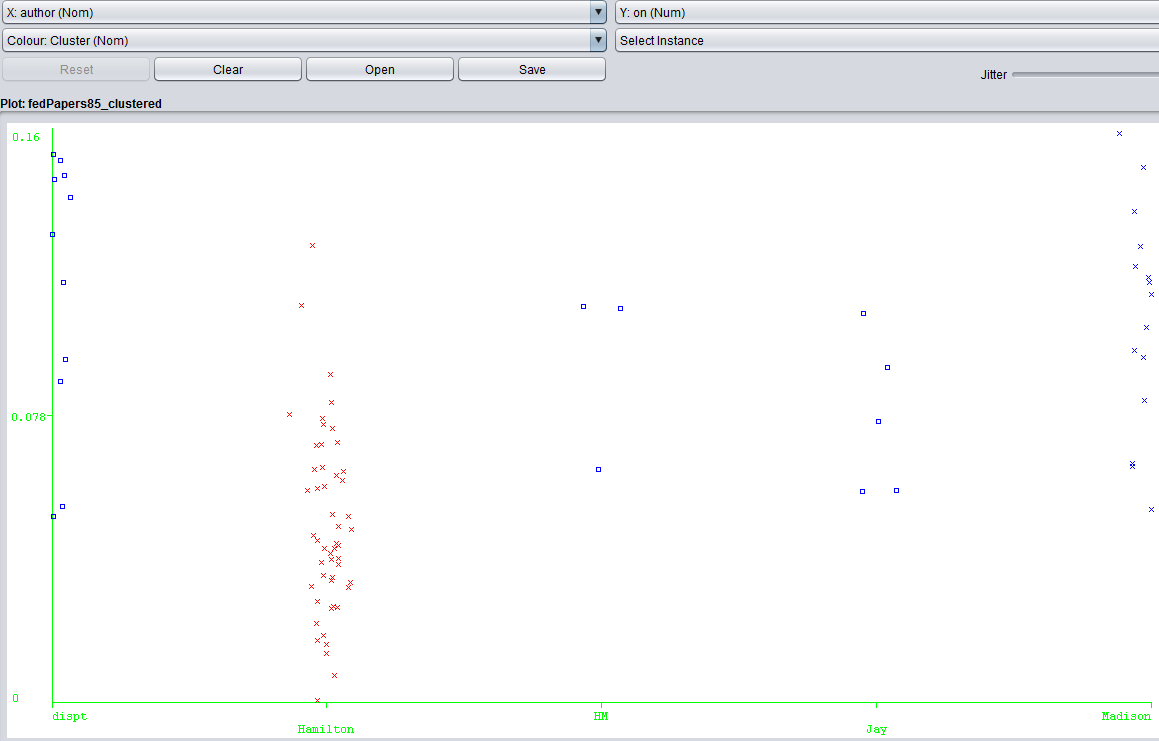


Using the EM Clusterer, we can see additional words that differentiate the two linguistic stylings (Looking at difference between the means of the clusters by attribute):

Such as ‘in’



‘on’ :



And ‘would’ :

