**Syracuse University**

**IST-736 Assignment 3**

ThulasiRam RuppaKrishnan

IST 736

Professor Ami Gates

Contents

[Introduction 3](#_Toc14651878)

[Analysis and Models 5](#_Toc14651879)

[**About the data** 5](#_Toc14651880)

[**Models** 9](#_Toc14651881)

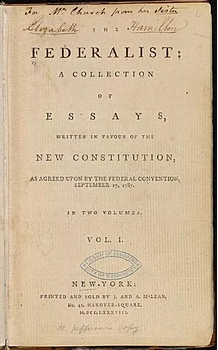
[Results 16](#_Toc14651882)

[Conclusion 18](#_Toc14651883)

## 

## **Introduction**

The Federalist (later known as The Federalist Papers) is a collection of 85 articles and essays written by Alexander Hamilton, James Madison, and John Jay under the pseudonym "Publius" to promote the ratification of the United States Constitution. The first 77 of these essays were published serially in the Independent Journal, the New York Packet, and The Daily Advertiser between October 1787 and April 1788. A two-volume compilation of these 77 essays and eight others was published as "The Federalist: A Collection of Essays", Written in Favor of the New Constitution, as Agreed upon by the Federal Convention, September 17, 1787, by publishing firm J. & A. McLean in March and May 1788.



The authors of The Federalist intended to influence the voters to ratify the Constitution. In "Federalist No. 1", they explicitly set that debate in broad political terms:

It has been frequently remarked, that it seems to have been reserved to the people of this country, by their conduct and example, to decide the important question, whether societies of men are capable or not, of establishing good government from reflection and choice, or whether they are forever destined to depend, for their political constitutions, on accident and force.

"Federalist No. 10" is generally regarded as the most important of the 85 articles from a philosophical perspective. In it, Madison discusses the means of preventing rule by majority faction and advocates a large, commercial republic. This is complemented by "Federalist No. 14", in which Madison takes the measure of the United States, declares it appropriate for an extended republic, and concludes with a great defense of the constitutional and political creativity of the Federal Convention. In "Federalist No. 84", Hamilton makes the case that there is no need to amend the Constitution by adding a Bill of Rights, insisting that the various provisions in the proposed Constitution protecting liberty amount to a "bill of rights." "Federalist No. 78", also written by Hamilton, lays the groundwork for the doctrine of judicial review by federal courts of federal legislation or executive acts. "Federalist No. 70" presents Hamilton's case for a one-person chief executive. In "Federalist No. 39", Madison presents the clearest exposition of what has come to be called "Federalism." In "Federalist No. 51", Madison distills arguments for checks and balances in an essay often quoted for its justification of government as "the greatest of all reflections on human nature."

According to historian Richard B. Morris, the essays that make up The Federalist Papers are an "incomparable exposition of the Constitution, a classic in political science unsurpassed in both breadth and depth by the product of any later American writer."

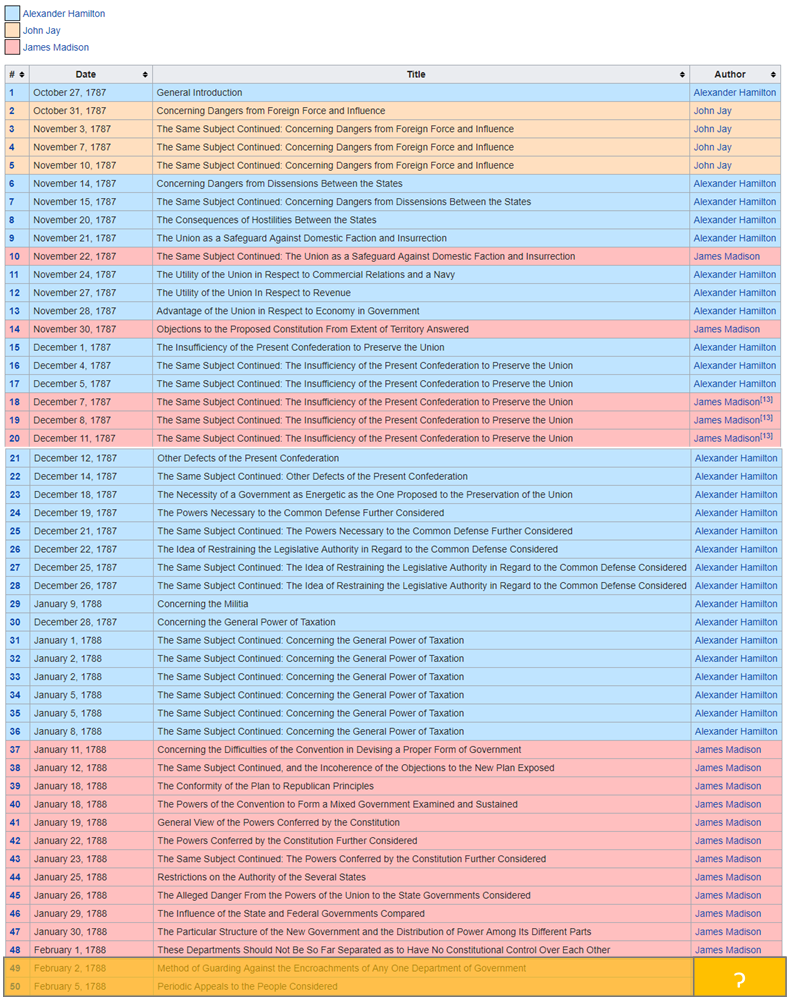
The authorship of certain of the Federalist essays was disputed from the beginning. Both Hamilton and Madison produced lists that claimed some of the same papers. There followed a series of lists, some claiming authorship for Madison and some for Hamilton.

The consensus of traditional scholarship, seconded by Mosteller and Wallace, allocates the papers: Hamilton 51 (1, 6-9, 11-13, 15-17, 21-36, 59-61, 65-85); Madison 29 (10, 14, 18-20, 37-58, 62, 63); Jay 5 (2-5, 64). Mosteller and Wallace set the boundary conditions for the subsequent non-traditional work – e.g., not using the Jay articles as a control. Most of these later practitioners do not select or prepare the input text as carefully as Mosteller and Wallace – and their selection and preparation were not as rigorous and complete as it should have been – as we will see.

## **Analysis and Models**

### **About the data**

The list of federalist papers is shown in **Table 1.1** and are highlighted with colors for each author.



****

**Table 1.1 Federalist papers by Author**

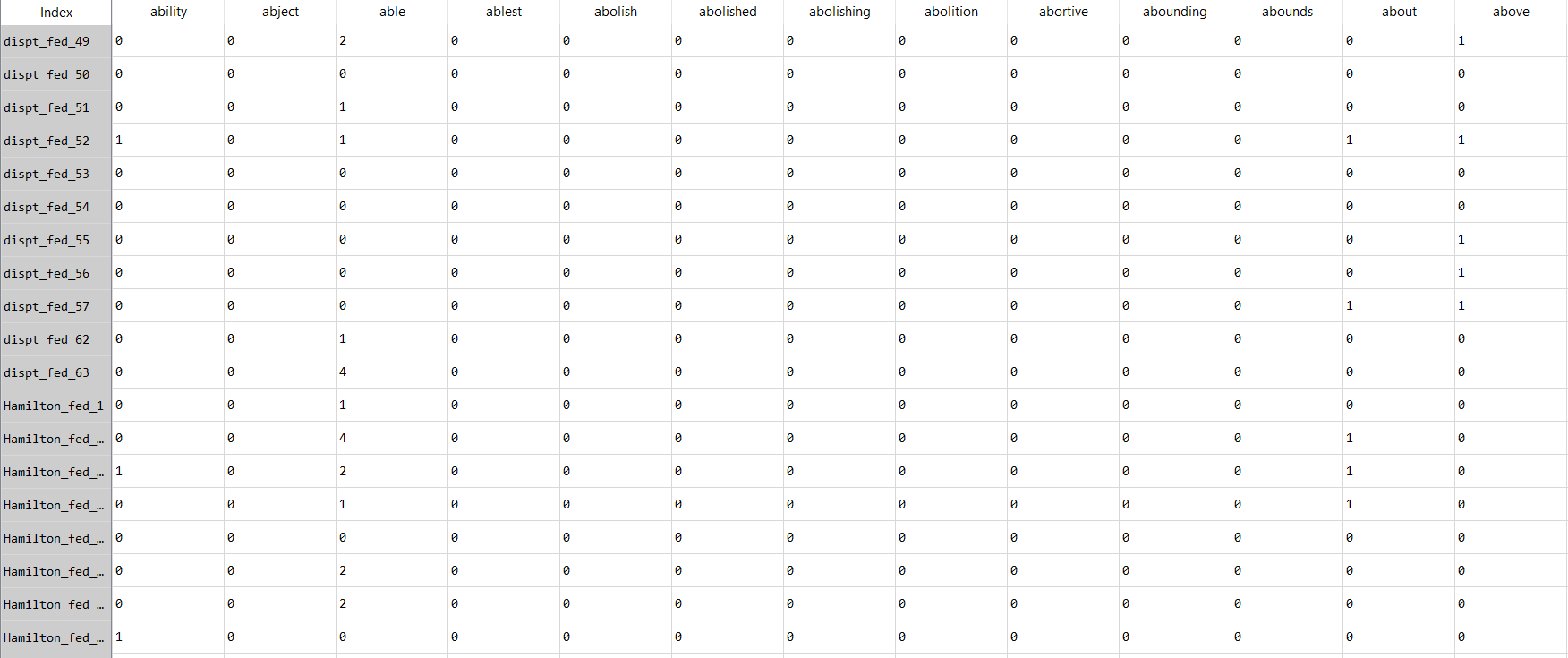
In the above list of papers, the ones which are highlighted is disputed between two authors Madison and Hamilton. Clustering techniques are applied here to solve the puzzle.

**Imported corpus and bag of words**



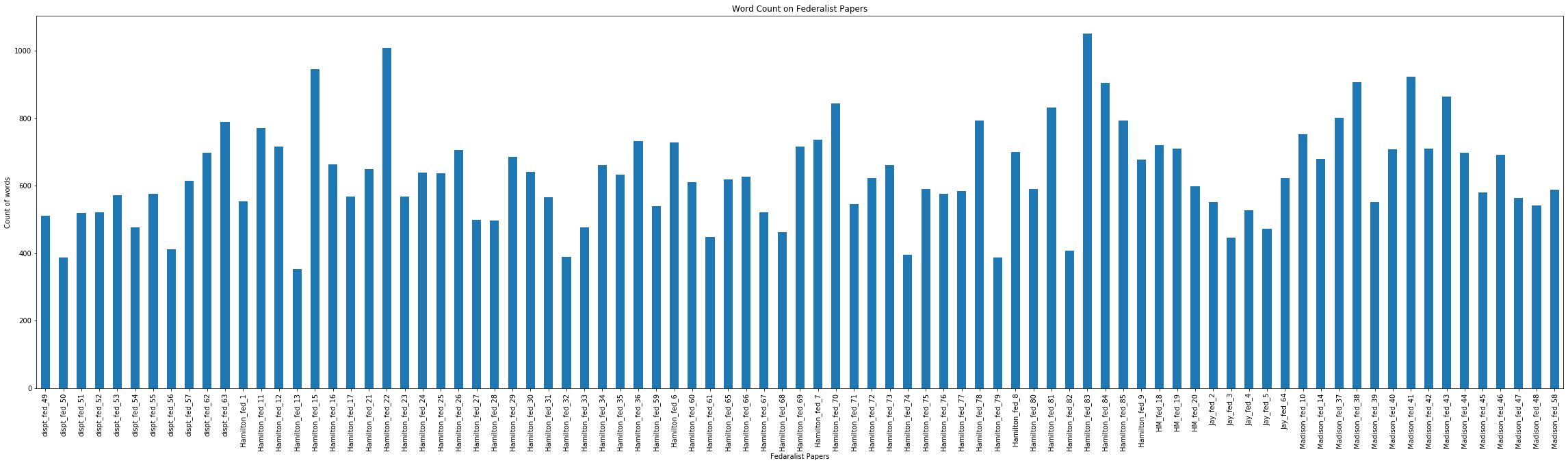
**Table 1.2 Papers**

**Corpus Data Frame**



**Table 1.3 Corpus Data Frame**

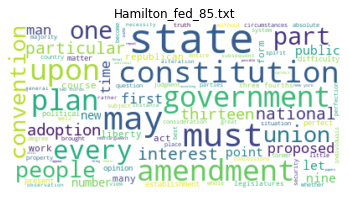
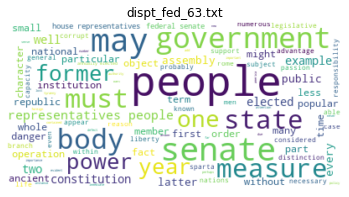
**Word count by papers**



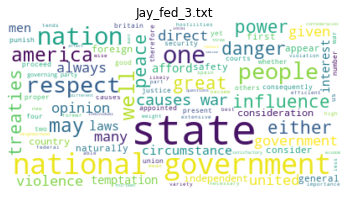
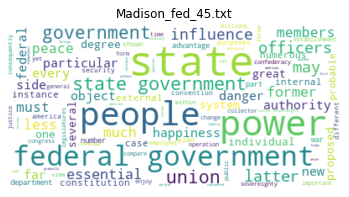
**Fig 1.1 Word count by papers**

**Word Clouds**

Fig 1.2,1.3,1.4 and 1.5 shows the word cloud of federalist papers by authors



**Fig 1.2 Word Cloud on Disputed Papers Fig 1.3 Word Cloud on Hamilton Papers**

**Fig 1.4 Word Cloud on Jay Papers Fig 1.5 Word Cloud of Madison Papers**

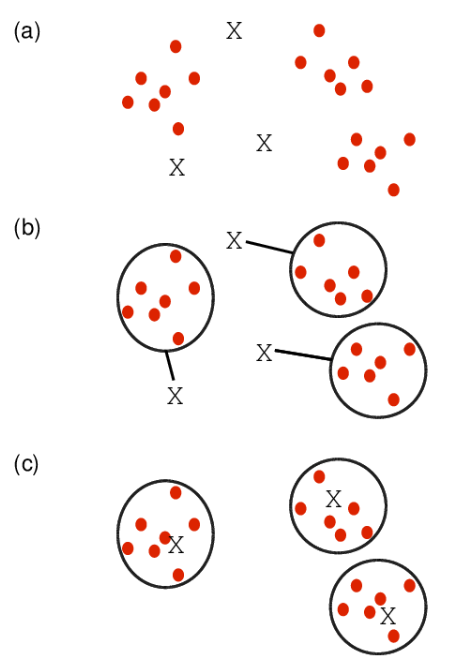
### **Models**

Clustering is the classification of data objects into similarity groups (clusters) according to a defined distance measure. It is used in many fields, such as machine learning, data mining, pattern recognition, image analysis, genomics, systems biology, etc. Machine learning typically regards data clustering as a form of unsupervised learning.

#### **Clustering Technique**

**K means clustering**

Unlike hierarchical clustering, K means require a number of clusters (k) as an input. This algorithm then randomly assigns items to the k clusters. Calculate new centroid for each of the k clusters and the distance of all items to the k centroids. Then assign items to the closest centroid. Repeat this process until clusters assignments are stable as illustrated in **Fig 2.2**

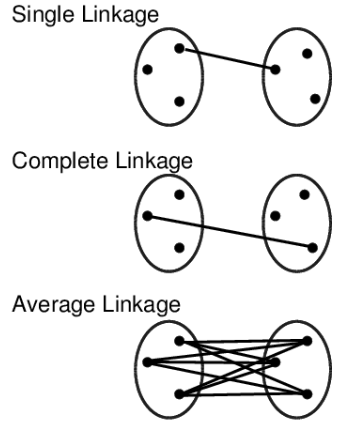


**Fig 2.2 K Mean Clustering**

*Source* [*http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html*](http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html)

#### **Cluster Linkage**

Distance between two clusters can be derived by using single, complete or average linkage methods as shown in **Fig 2.3**. Single linkage uses the minimum distance between two points whereas the complete linkage used the maximum distance between two points. Average linkage calculates an average of the distance between all points between the clusters



**Fig 2.3 Cluster Linkage**

*Source* [*http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html*](http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html)

In this exercise, tools used to vectorize the documents are as follows

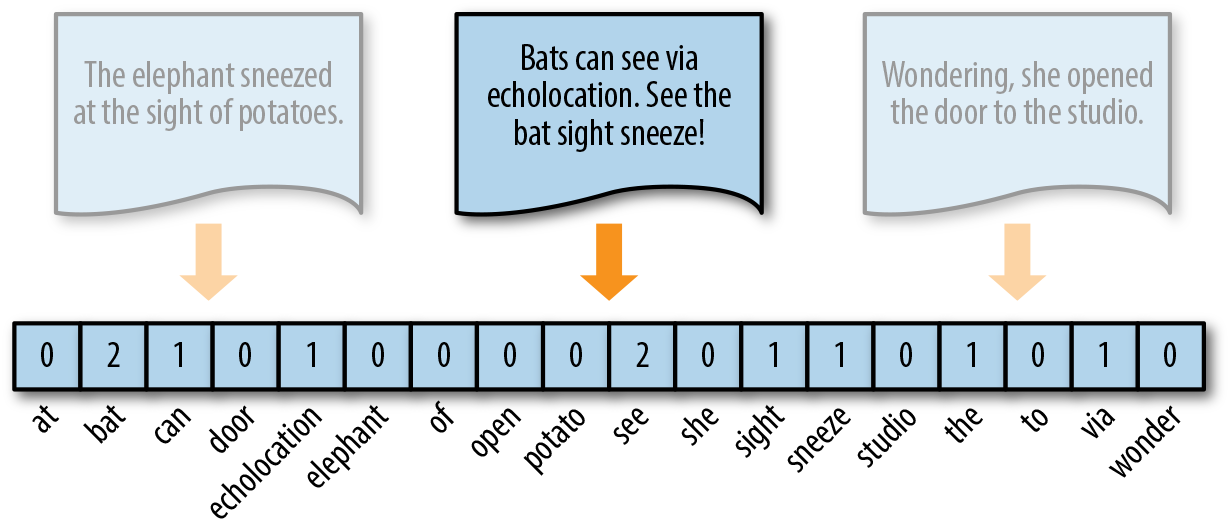
* NLTK
* Sklearn

Vector encoding methods that are used for this analysis with the help of using NLTK and Sklearn packages are as follows

* **Frequency**
* **Term Frequency**
* **TF-IDF**

#### **Frequency Encoding**

The simplest vector encoding model is to simply fill in the vector with the frequency of each word as it appears in the document. In this encoding scheme, each document is represented as the multiset of the tokens that compose it and the value for each word position in the vector is its count. This representation can either be a straight count (integer) encoding as shown in **Figure 2.1** or a normalized encoding where each word is weighted by the total number of words in the document.



**Figure 2.1 Frequency Encoding**

Vectors can become extremely sparse, particularly as vocabularies get larger, which can have a significant impact on the speed and performance of machine learning models. For very large corpora, it is recommended to use the Scikit-Learn HashingVectorizer, which uses a hashing trick to find the token string name to feature index mapping. This means it uses very low memory and scales to large datasets as it does not need to store the entire vocabulary and it is faster to pickle and fit since there is no state. However, there is no inverse transform (from vector to text), there can be collisions, and there is no inverse document frequency weighting.

#### **TF Encoding**

Term Frequency is defined as how frequently the word appear in the document or corpus. As each sentence is not the same length so it may be possible a word appears in long sentence occur more time as compared to word appear in sorter sentence. Term frequency can be defined as:

**tf(t,d) = count of t in d / number of words in d**

#### **TF-IDF Encoding**

The bag-of-words representations that we have explored so far only describe a document in a standalone fashion, not considering the context of the corpus. A better approach would be to consider the relative frequency or rareness of tokens in the document against their frequency in other documents. The central insight is that meaning is most likely encoded in the rarer terms from a document. For example, in a corpus of sports text, tokens such as “umpire,” “base,” and “dugout” appear more frequently in documents that discuss baseball, while other tokens that appear frequently throughout the corpus, like “run,” “score,” and “play,” are less important.

Document Frequency measures the importance of document in whole set of corpus, this is very similar to TF. The only difference is that TF is frequency counter for a term t in document d, whereas DF is the count of occurrences of term t in the document set N. In other words, DF is the number of documents in which the word is present. We consider one occurrence if the term consists in the document at least once, we do not need to know the number of times the term is present.

**df(t) = occurrence of t in documents**

To keep this also in a range, we normalize by dividing with the total number of documents. To know the informativeness of a term, and DF is the exact inverse of it. Inverse Document Frequency (IDF) is the inverse of the document frequency which measures the informativeness of term t. It will be very low for the most occurring words such as stop words (because stop words such as “is” is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

**idf(t) = N/df**

Now there are few other problems with the IDF, in case of a large corpus, say 10,000, the IDF value explodes. So, to dampen the effect we take log of IDF.

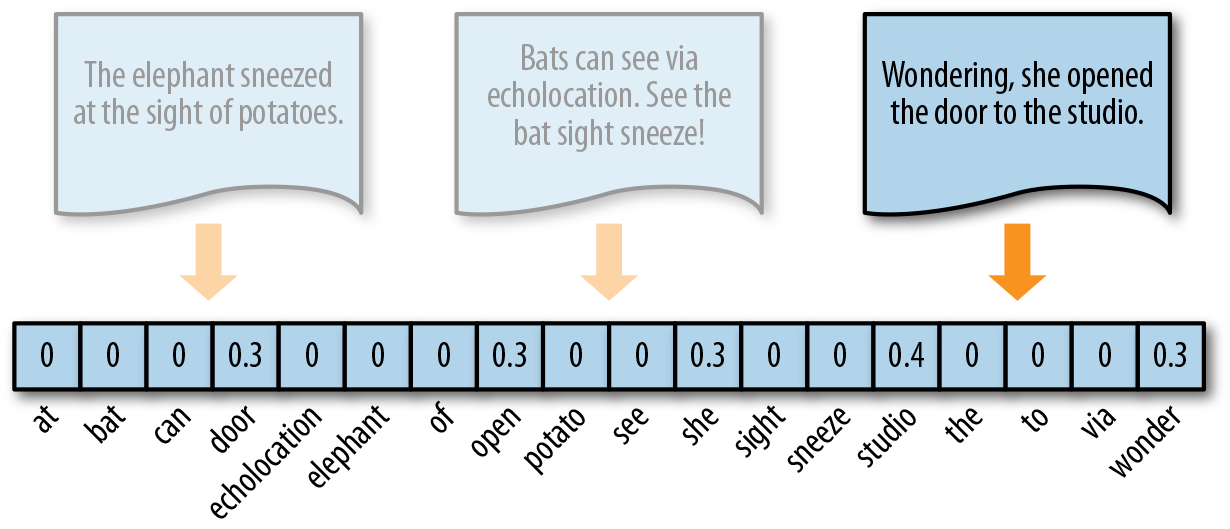
During the query time, when a word which is not in vocab occurs, the df will be 0. As we cannot divide by 0, we smoothen the value by adding 1 to the denominator.

**idf(t) = log (N/(df + 1))**

Finally, by taking a multiplicative value of TF and IDF, TF-IDF score is measured.

**tf-idf(t, d) = tf(t, d) \* log(N/(df + 1))**

As shown in **Figure 2.2**, where the token studio has a higher relevance to this document since it only appears there.



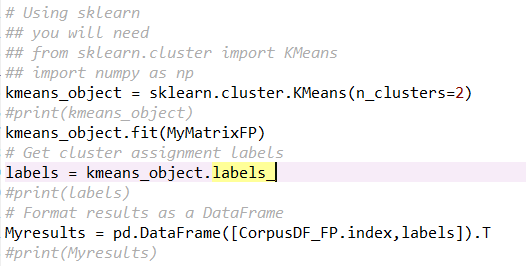
**Figure 2.2 TF-IDF Encoding**

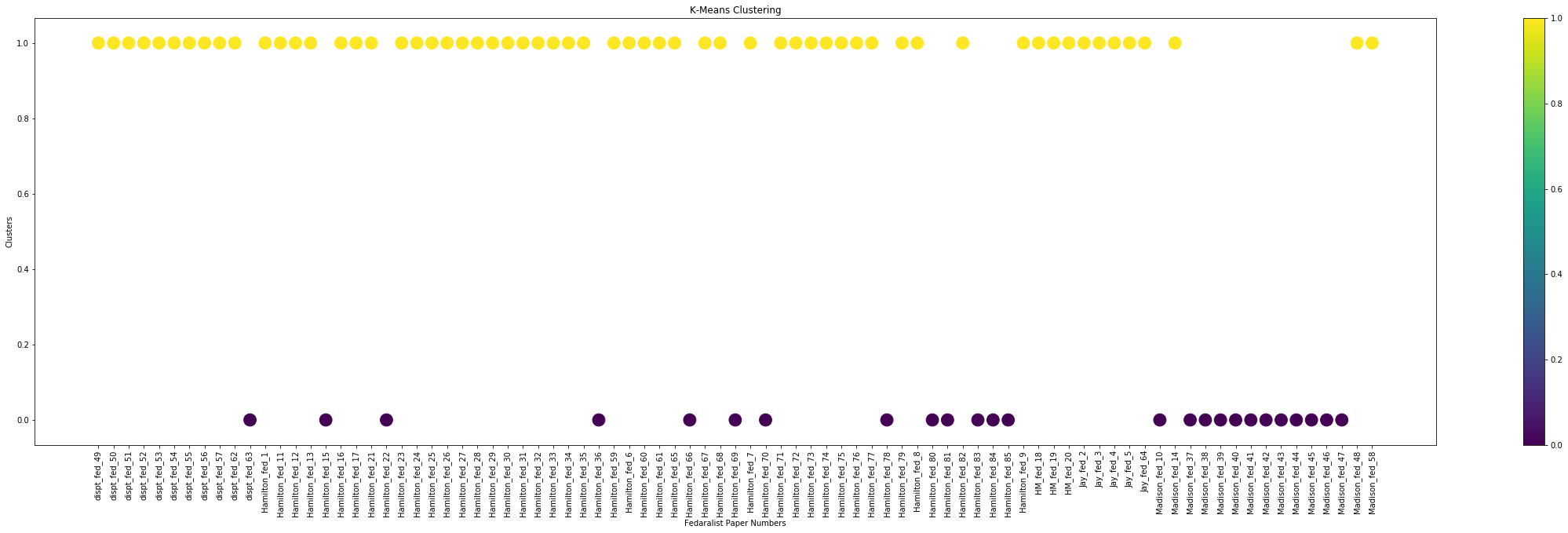
Source <https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089>

#### **Model 1: K Means Clustering with 2 clusters**

**Model specification 1: No transformation and Vectorization using frequency count**

K means models with 2 clusters are modelled on all 85 papers and the results of the clusters are shown in **Figure 2.3**

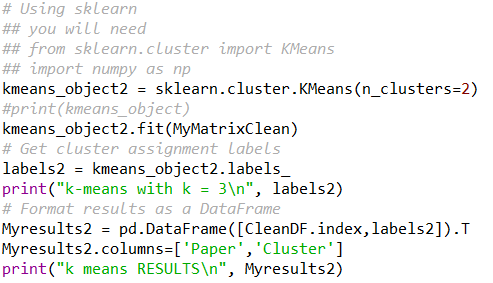


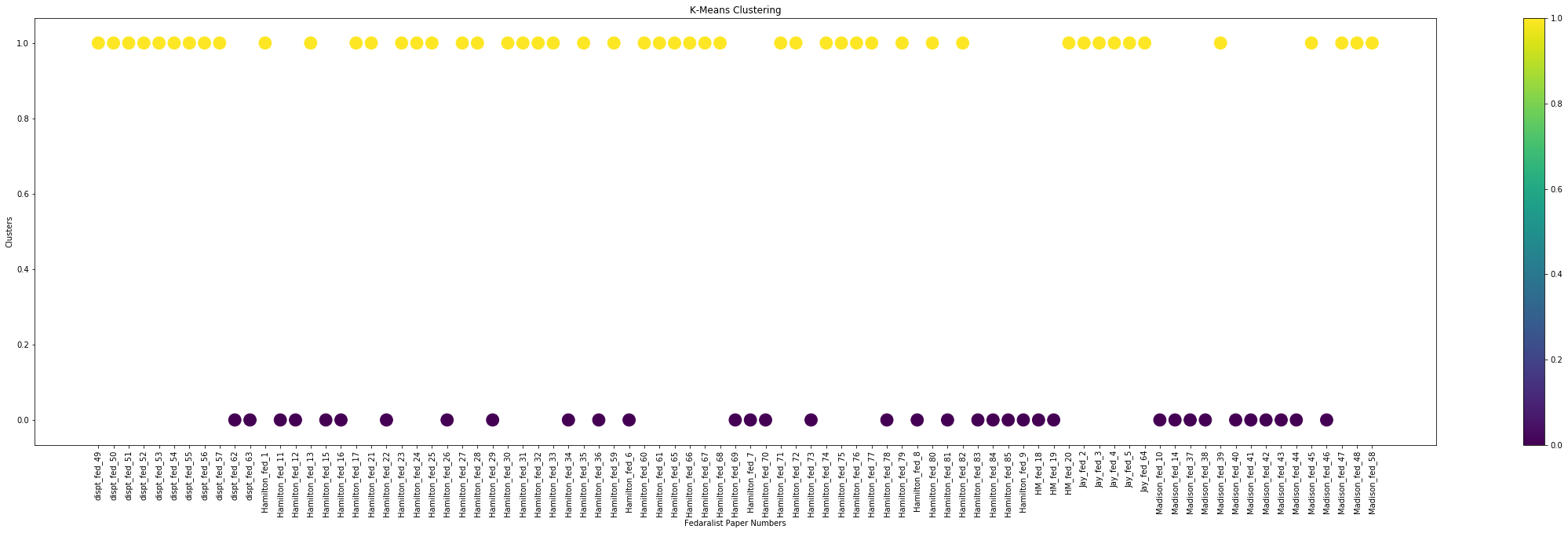


**Figure 2.3 K means clusters with 2 Clusters**

**Model specification 2: StopWords removed, Lemmatized, numbers are ignored and vectorized using frequency count**

K means models with 2 clusters for all 85 papers are applied on top of tokens that are cleaned by removing stopwords, numbers and lemmatized. Results of the clusters are shown in **Figure 2.4**

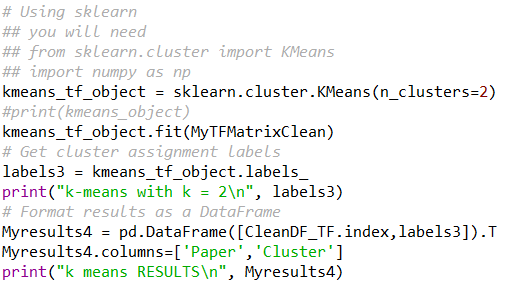


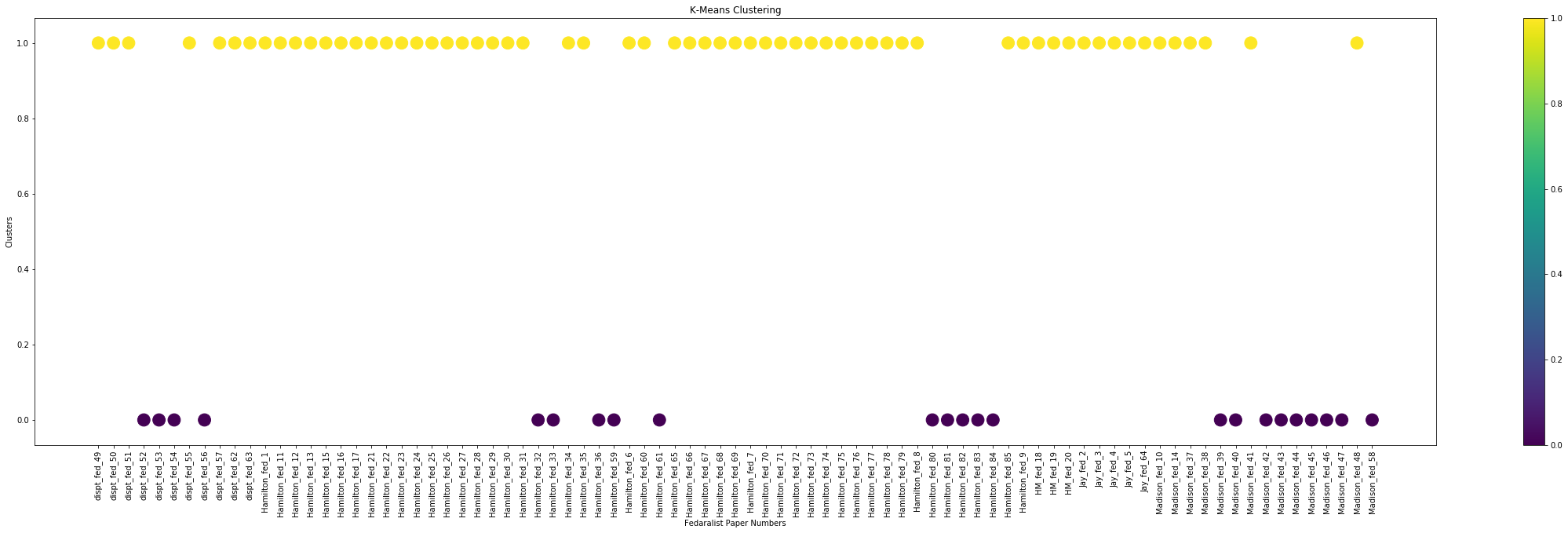


**Figure 2.4 K Means cluster on Frequency count**

**Model specification 3: StopWords removed, Lemmatized, numbers are ignored and vectorized using Term Frequencies**

K means models with 2 clusters for all 85 papers are applied using term frequencies on top of tokens that are cleaned by removing stopwords, numbers and lemmatized. Results of the clusters are shown in **Figure 2.5**

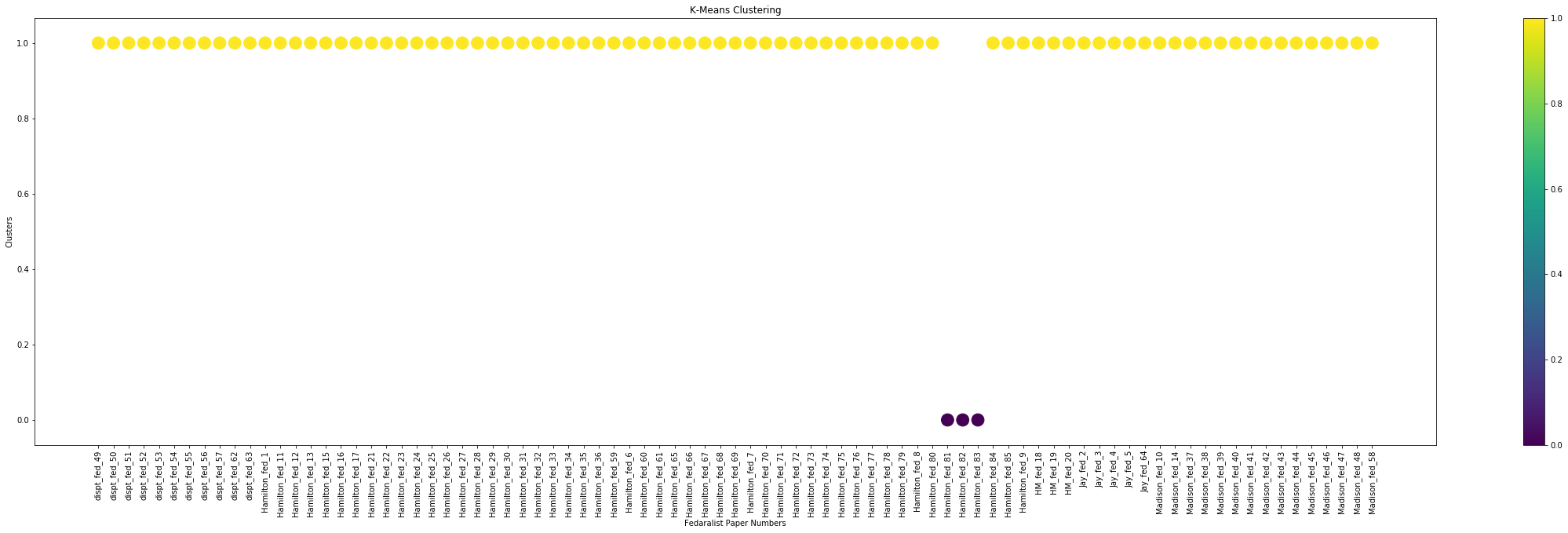




**Figure 2.5 K Means cluster on Term Frequencies**

**Model specification 4: StopWords removed, Lemmatized, numbers are ignored and vectorized using TFIDF**

K means models with 2 clusters for all 85 papers are applied using TFIDF on top of tokens that are cleaned by removing stopwords, numbers and lemmatized. Results of the clusters are shown in **Figure 2.6**



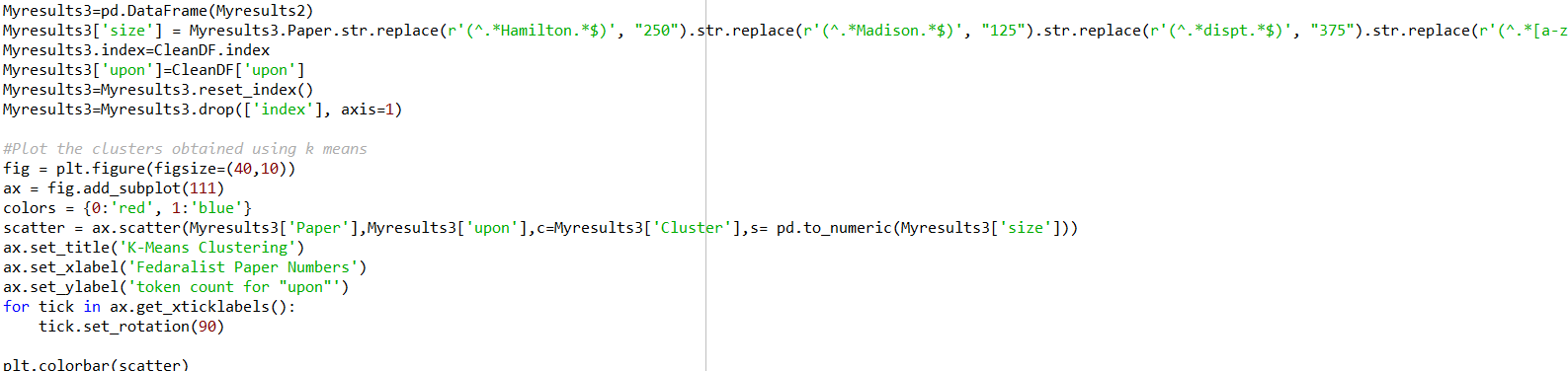
**Figure 2.6 K Means cluster on TFIDF**

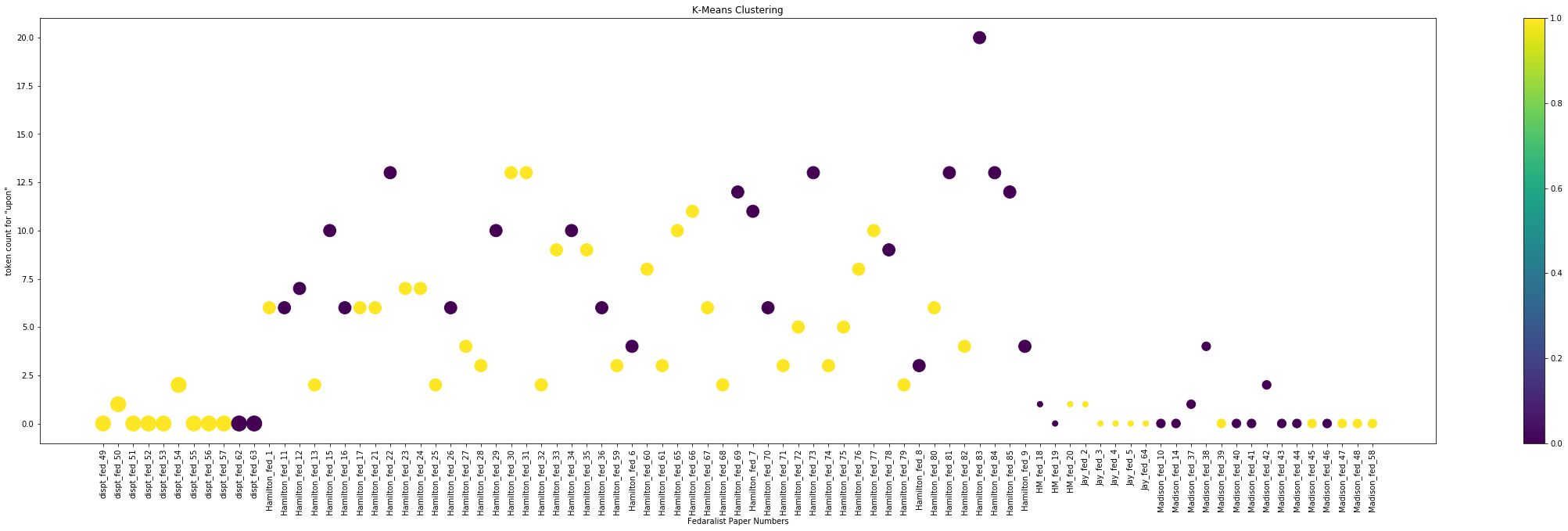
## **Results**

#### **Model 1: K Means Clustering with 2 clusters**

**Results for Model specification 2: StopWords removed, Lemmatized, numbers are ignored and vectorized using frequency count**

Most of the disputed papers are clustered with Hamilton but there is a clear line of difference between Hamilton and disputed papers with respect to the count on the token “upon” **Figure 3.1**

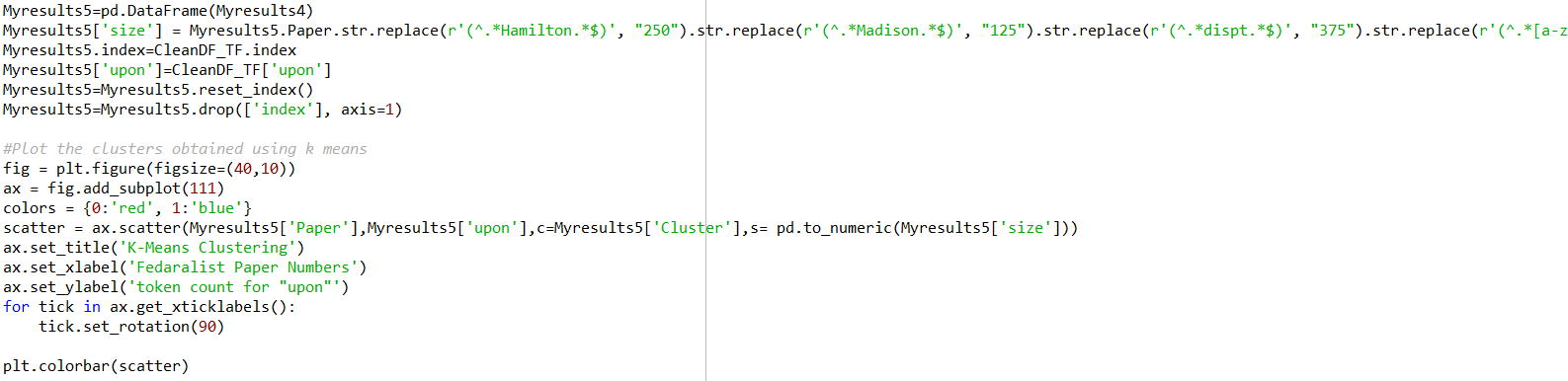


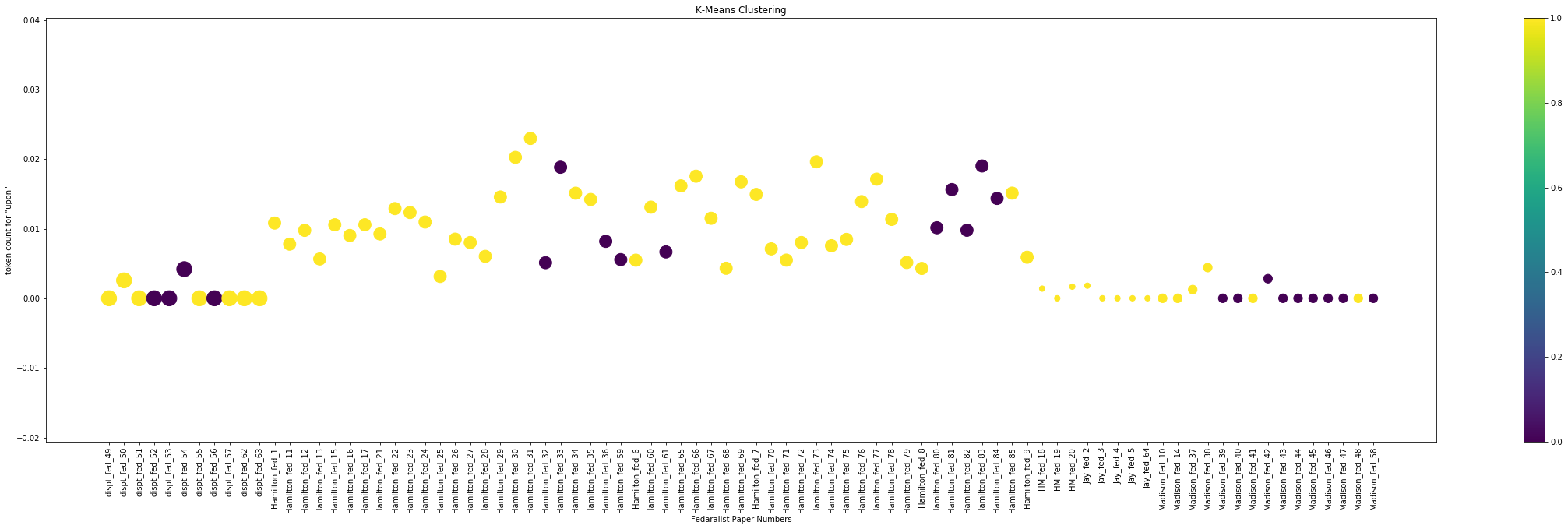


**Figure 3.1 K Means cluster on Frequency count and token “upon” count**

**Results for Model specification 3: StopWords removed, Lemmatized, numbers are ignored and vectorized using Term Frequencies**

Most of the disputed papers are clustered with Hamilton but there is a clear line of difference became smaller between Hamilton and disputed papers when compared to Frequency count plot **Figure 3.2**



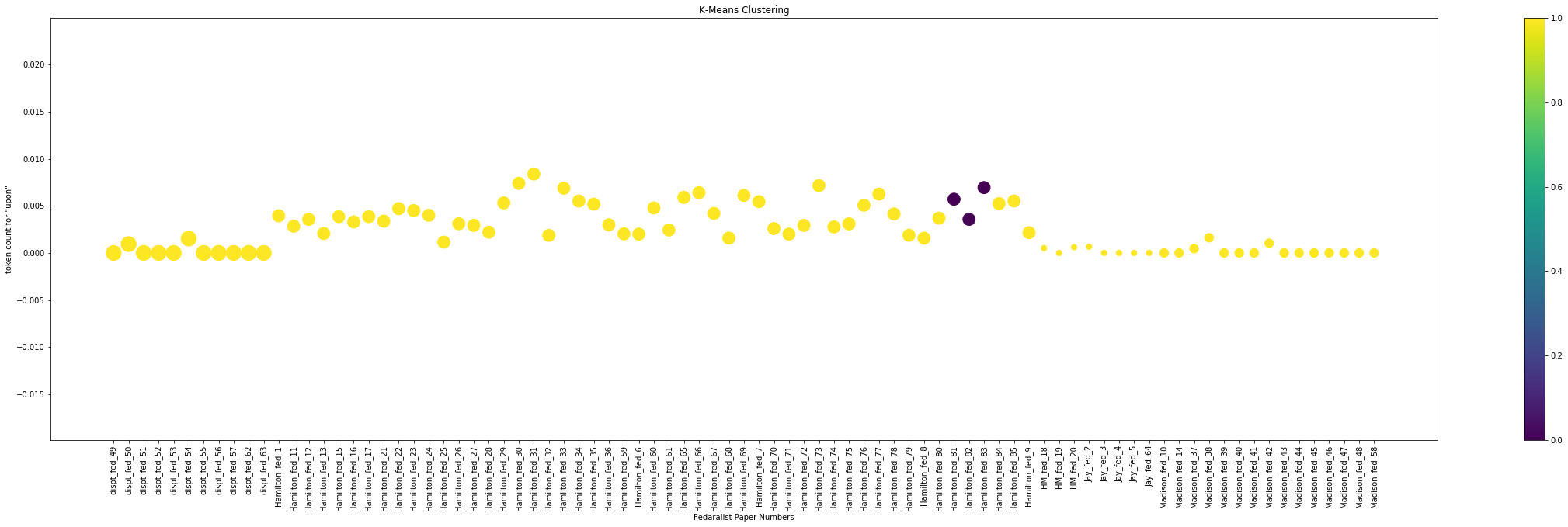


**Figure 3.2 K Means cluster on Term Frequencies and the TF for token “upon”**

**Results for Model specification 4: StopWords removed, Lemmatized, numbers are ignored and vectorized using TFIDF**

Almost all the papers are clustered as one except 3. All the disputed papers are clustered with Hamilton but there is a clear line of difference became least between Hamilton and disputed papers with respect to the TFIDF on the token “upon” **Figure 3.3**

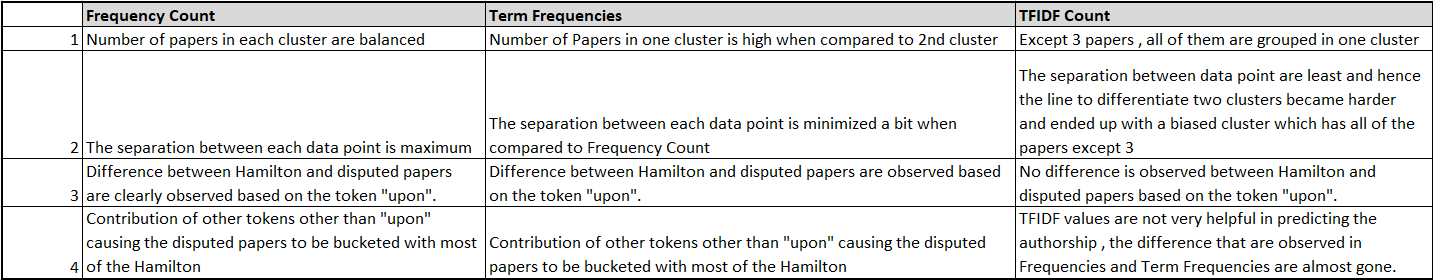




**Figure 3.3 K Means cluster on Term Frequencies and the TFIDF for token “upon”**

## **Conclusion**

After analyzing the results of clusters obtained from different vectorization technique, it is observed that the TFIDF approach is less accurate and ended up with a cluster with all papers except 3. Other differences that are observed are explained in the below **Table 4.1 Observation on Different Vectorization methods**



**Table 4.1 Observation on Different Vectorization methods**

**Hamilton papers**

Hamilton papers are significant in numbers when compared with other authors. Most of the Hamilton papers exhibit the same pattern except for few that are distinct from the general pattern. This shows that Hamilton has got a similar writing pattern and the variation on his essays are not very large.

**Madison papers**

Madison papers are not very large in number like Hamilton. With fewer papers compared to Hamilton, Madison papers are falling into two to three patterns. This indicates that Madison had a very different writing pattern in each of his essay showing more considerable variations when compared with all of his essays.

**Dispute papers**

Dispute paper shows a similar pattern like Hamilton papers. All the dispute papers when compared with each other shows fewer variations and can be grouped with the writing style of Hamilton. It didn’t show more significant differences like Madison papers within its group. Also, the writing style of dispute papers are very different from Madison style and couldn’t group them under the same bucket of Madison papers. This suggests more inclination towards the authorship of Hamilton and not Madison