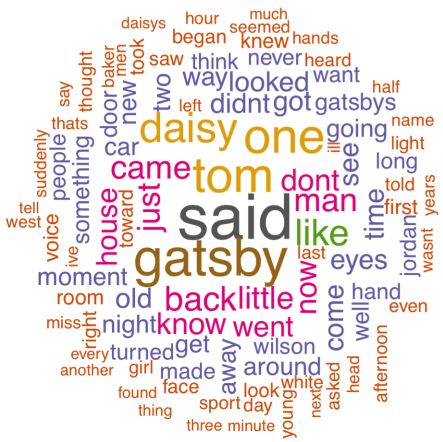
Research in the field of text mining suggests that distinct writing styles are discernible by word selection and frequency. Even the use of common words can help distinguish one writer from another. We will analyze the writing styles of four authors based on their novels written during 1915-1940 period. Table 1 displays the list of analyzed novels.

The text of each novel is first converted to a data frame and then to a corpus. For each novel, the frequencies of appeared words were counted. The corpus needs a couple of transformations, including changing letters to lower case, removing punctuations, numbers and stopwords (the, a, of, which, etc.). After building a matrix with the frequencies of the words, we showed the importance of the words with a word cloud where frequent words are plotted first, which makes them appear in the center of the cloud (Figure 1). Lists of twenty of the most-used words for two novels are shown in Figure 2. Based on the word clouds and lists of the most-used words for all 13 novels, we selected fifteen top words they share and calculate their frequencies for each book.

Author	Novel	Year
Somerset Maugham	Of Human Bondage	1916
Somerset Maugham	The Moon and Sixpence	1919
Somerset Maugham	The Painted Veil	1937
Somerset Maugham	Theatre	1937
Scott Fitzgerald	The Great Gatsby	1925
Scott Fitzgerald	Tender is the Night	1934
Jack London	The Jacket	1915
Jack London	The Little Lady of the Big House	1916
Jack London	Jerry of the Islands	1917
Jack London	Hearts of Three	1920
Virginia Woolf	Mrs Dalloway	1925
Virginia Woolf	To the Lighthouse	1927
Virginia Woolf	Orlando	1928

Table 1: List of novels

Using this data, we analyzed it using cluster analysis and principal components. Clusters of authors are shown in Figure 3. *Ward's* hierarchical clustering method, which builds clusters incrementally and computes the sum of square distances within each cluster, was used. At the beginning, each observation assigned to its own cluster, then at each iteration, the most similar two clusters are merged until all of the clusters have been merged. We expect to see four clusters, each corresponding to a different author. However, clustering dendrogram reveals about five clusters. In general, the works of each author are similar to each other, which indicates that each author has a unique writing style. But we can see that *Orlando* and *The Little Lady of the Big House* novels belong to the one cluster. It seems that they have a similar style even they were written by the two different authors. Also, Somerset Maughams earliest work *Of Human Bondage* and



(a) S.Fitzgerald 'The Great Gatsby'

(b) V.Woolf 'Mrs. Dalloway'

Figure 1: Word clouds

word	freq	prop
strickland	404	0.007100550
said	312	0.005483593
know	218	0.003831485
one	217	0.003813909
see	173	0.003040582
made	164	0.002882402
little	161	0.002829675
never	160	0.002812099
like	159	0.002794523
man	155	0.002724221
think	150	0.002636343
stroeve	147	0.002583616
thought	138	0.002425435
now	129	0.002267255
can	126	0.002214528
come	121	0.002126650
dont	121	0.002126650
asked	120	0.002109074
mrs	120	0.002109074
life	110	0.001933318

(a) S.Maugham 'The Moon and Sixpence'

word	freq	prop
francis	757	0.008655187
henry	468	0.005350895
torres	399	0.004561981
leoncia	353	0.004036039
one	319	0.003647298
will	306	0.003498662
man	283	0.003235691
back	262	0.002995587
said	251	0.002869818
old	235	0.002686881
now	208	0.002378176
time	193	0.002206673
two	184	0.002103771
like	183	0.002092337
men	183	0.002092337
eyes	181	0.002069470
know	180	0.002058037
hand	175	0.002000869
way	172	0.001966568
well	159	0.001817932

(b) J.London 'Hearts of three'

Figure 2: Lists of twenty of the most-used words

Author	Novel	Year	cluster
Jack London	The Jacket	1915	1
Jack London	Jerry of the Islands	1917	1
Jack London	Hearts of Three	1920	1
Virginia Woolf	Orlando	1928	2
Jack London	The Little Lady of the Big House	1916	2
Somerset Maugham	Of Human Bondage	1916	3
Somerset Maugham	The Moon and Sixpence	1919	3
Somerset Maugham	The Painted Veil	1937	3
Somerset Maugham	Theatre	1937	3
Scott Fitzgerald	The Great Gatsby	1925	3
Scott Fitzgerald	Tender is the Night	1934	3
Virginia Woolf	Mrs Dalloway	1925	4
Virginia Woolf	To the Lighthouse	1927	4

Table 2: k-means clustering results

Scott Fitzgeralds novels have similar writing styles.

We applied *k-means* ($k = 4$) clustering to see if the novels by the same authors are clustered together. The algorithm begins by creating k centroids, then each iteration it assigns each data point to its closest centroid and calculates the new means of the observations in the new cluster. The results of the k-means clustering are shown in Table 2 (70% accuracy). Although *k-means* was able to cluster the works by the same author together, it didn't distinguish between Fitzerald's and Maugham's novels.

For further analysis, we applied a *principal component analysis (PCA)*, using certain word frequencies as variables. *PCA* is a dimension-reduction technique that can be used to reduce a large set of variables to a smaller set that still preserves most of the information in the large set. It seeks for the sets of correlated variables and transforms them into a smaller number of uncorrelated variables called principal components. *PCA* is searching for the linear combinations of variables such that the maximum variance is extracted from these variables.

Let's plot the first two principal components with the names of the authors and their works (see Figure 4). There do appear to be different clusters of authors, perhaps 3-5 clusters. It appears that Jack London's works are clustered close to each other. Fitzgerald's novel *Tender is the Night* is clustered close to Jack London's works. Other Fitzgerald's novel *The Great Gatsby* is clustered close to Somerset Maugham's works. Two works of Virginia Woolf are in the one cluster, but the last one forms its own cluster.

As we have seen, the clustering methods didn't cluster novels perfectly. We think that the main reason of such behavior is the word selection process being used. It is simply not enough to use 15 top words for the accurate analysis. Further, we apply more

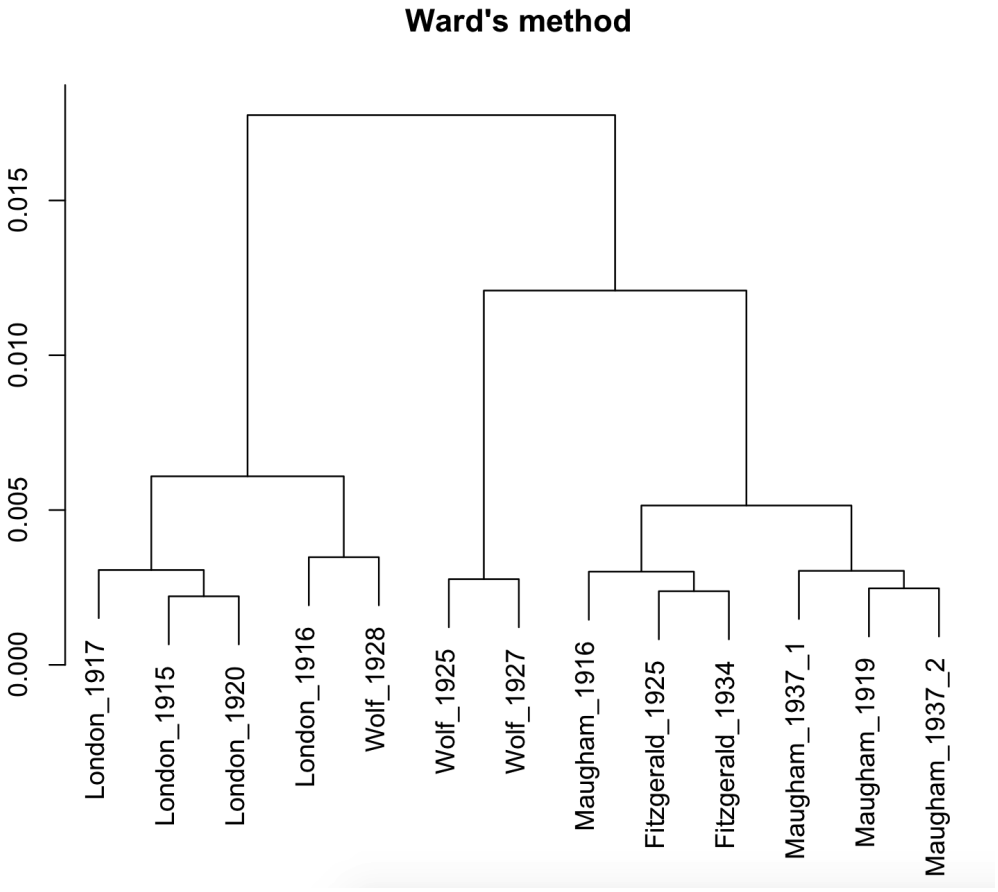


Figure 3: Cluster Dendrograms

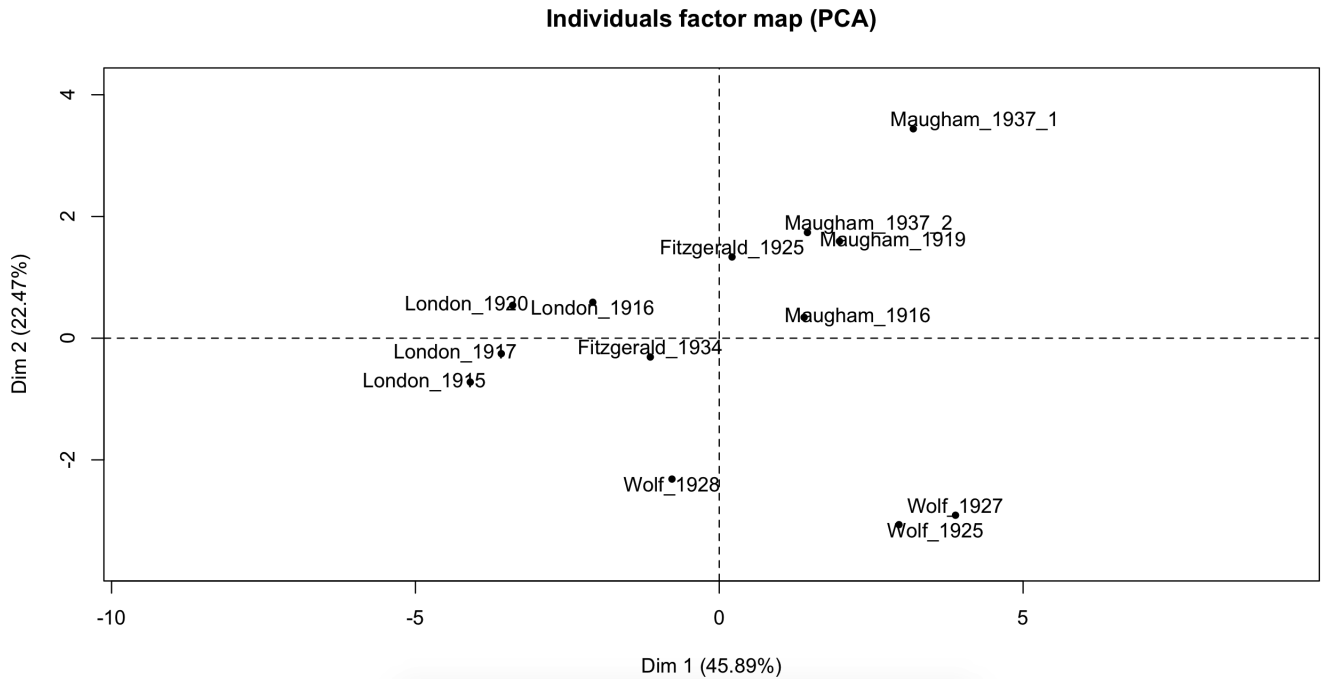


Figure 4: First two principal components for the data in Table 1

sophisticated approaches for more accurate clustering.

We will focus our analysis on only London's three novels and Maugham's two novels, and show how logistic regression and support vector machine methods perform on distinguishing between these two authors writing styles. To increase our sample, we split each novel into chapters. There are 87 chapters corresponding to Maughams novels and 82 chapters corresponding to Londons works.

After that, we construct a *TF-IDF* (term frequency-inverse document frequency) matrix that describes the relative frequency of the words in the document against their frequency in other documents. This approach reduces the weight given to common words and highlights the rare words in a document. We believe that *TF-IDF* can achieve better results than simple word frequencies.

We then applied *k-means* on the tf-idf matrix, and got 86.98% of clustering accuracy (22 out of 169 chapters were misclassified). Since the matrix is sparse and has 15704 columns, each corresponding to the certain word, we reduce its dimension by applying *PCA* to each row of the matrix corresponding to the certain novel's chapter. Figure 5 displays two principal components which explain only 12% of the total variance. We dont observe well-separated clusters, it seems that some chapters belonging to the different authors mixed with each other.

We split our data into training and testing set (80% and 20%). We didn't consider a

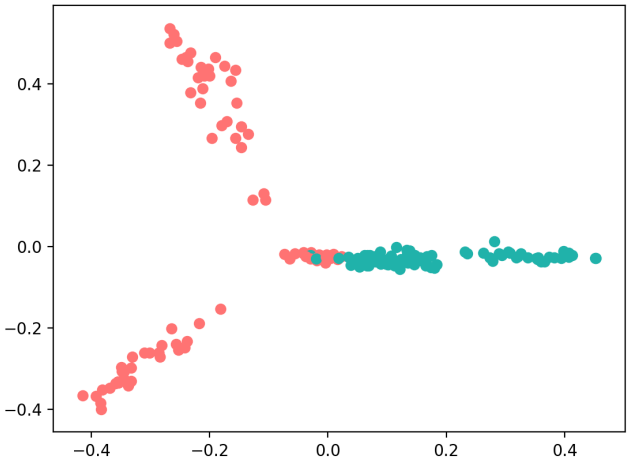


Figure 5: First two principal components of the tf-idf matrix

Method	number of principal components	train score	test score
LR	2	91.85	88.24
LR	3	91.85	91.18
LR	4	100.0	100.0
SVM	2	88.89	88.24
SVM	3	90.37	88.24
SVM	4	100.0	100.0

Table 3: Supervised learning classification results

validation set, since our sample is not large enough. Then we applied *logistic regression (LR)* and *support vector machine (SVM)* approaches on the first two, three, and four principal components, respectively. At the end, keeping only the first four principal components resulted in high accuracy.

As we have seen, using the first few principal components might result in high classification accuracy. In this analysis, we were able to reduce our matrix dimension from $(169, 15704)$ to $(169, 4)$.

Lastly, let's consider a *cosine˙similarity* metric used to determine how similar two documents are irrespective of their size. *Cosine similarity* calculates similarity by measuring the cosine of the angle between two vectors. Since we have already constructed a matrix of *TF-IDF* scores, we can consider its rows as the input vectors. We calculated *cosine˙similarity* values for few chapters corresponding to the different novels to find their top 10 related chapters. As expected, all of these chapters were similar to the chapters of the same novel. Computing such metrics as *cosine˙similarity* metric helps to answer a

question of how similar different documents are to each other.