Clustering Assignment

Group 2

Supervised by: DR. Arya Rahgozar

Choosing data:

First we did some web scrapping to get 5 books which are (the war of the worlds),

(Emma), (the King James Version of the bible), (paradise lost), and (the tragedy of hamlet, prince of Denmark)

We wanted to have a variety of topics within the different genres.

Preprocessing and Data Cleansing:

Then we started with preprocessing as we removed the special characters, stop words and the unusual spaces as cleaning for the data.

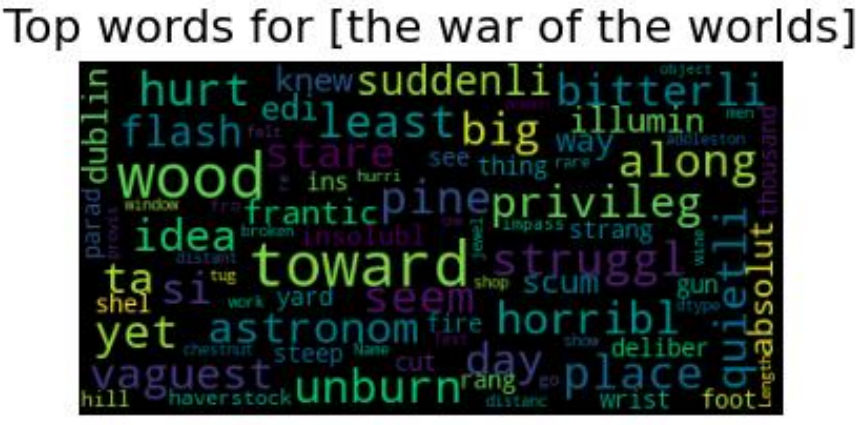
Now we have our data prepared so we grouped our data by the name of the book and we have encoded the labels using LabelEncoder.

To be able to deal with our data we transformed it BOW and TF-IDF and LDA.

To reduce the number steps in our code and to have a better shape for the data, we used the pipeline technique.

To visualize our data we did **EDA** "Exploring data analysis", by visualizing the most repeated 100 word for each class.

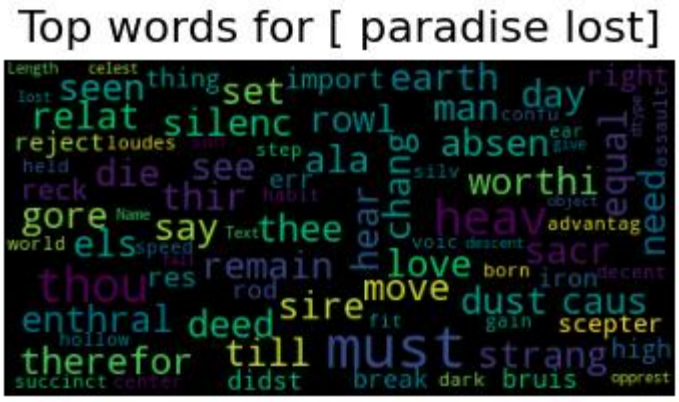
1) Top words for the war of the worlds



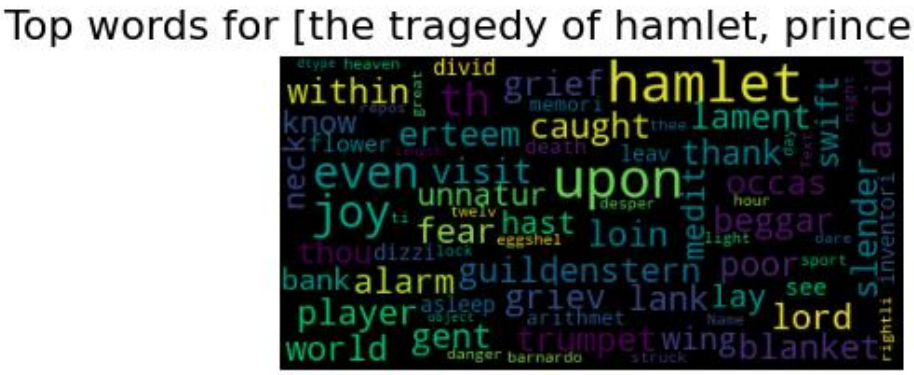
2) Top words for Emma

3) Top words for the King James Version of the bible

4) Top words for paradise lost



5) Top words for the tragedy of hamlet, prince of Denmark



Feature Engineering:

We have used pipeline technique to make our code easier to be manipulated and then we have tested them with multiple combination of feature engineering techniques like “BOW”, “TF-IDF”, “LDA”, “PCA” on different models so we can determine the best feature engineering combination to contribute in raising our accuracy the most.

Using k-means, hierarchical, and Gaussian matrix:

We have used three clustering techniques which are: (k-means, hierarchical, and Gaussian matrix).

We used each feature engineering technique with the clustering techniques using “pipeline”

For k-means pipeline:

We have trained multiple feature engineering and we have got multiple elbow results, we have noticed that when the result is near to 5 it gave us the best accuracy that was found in k-means with tf-idf.

For hierarchical pipeline:

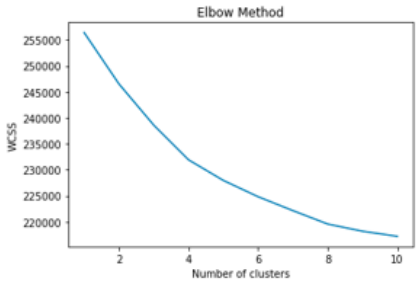
We have trained multiple feature engineering like what we have done with k-means, we have found that the hierarchical with BOW have the best accuracy.

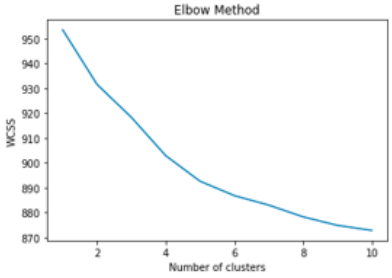
For Gaussian matrix:

We have trained multiple feature engineering and we have used PCA for BOW and TF-IDF models, because the number of features was huge so we used it to do some dimensional reduction and we have kept the LDA as it is.

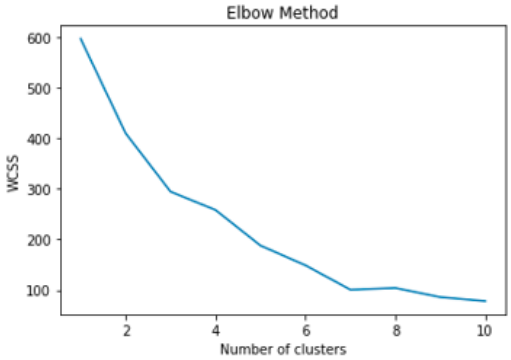
K-means

K-MEANS WITH BOW 65% V-Score

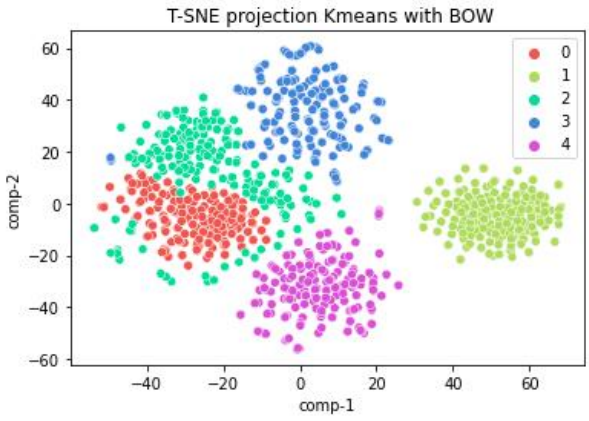
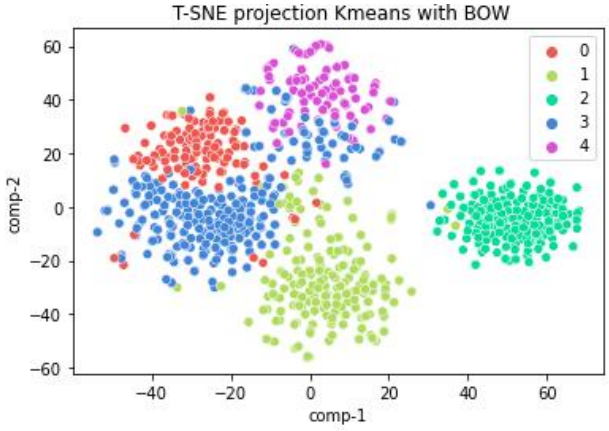


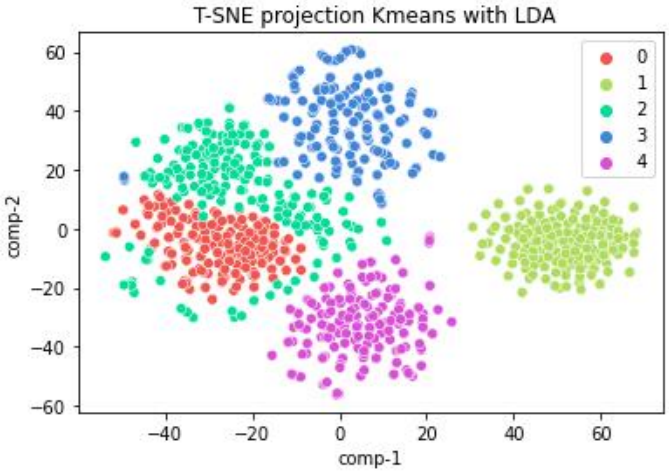


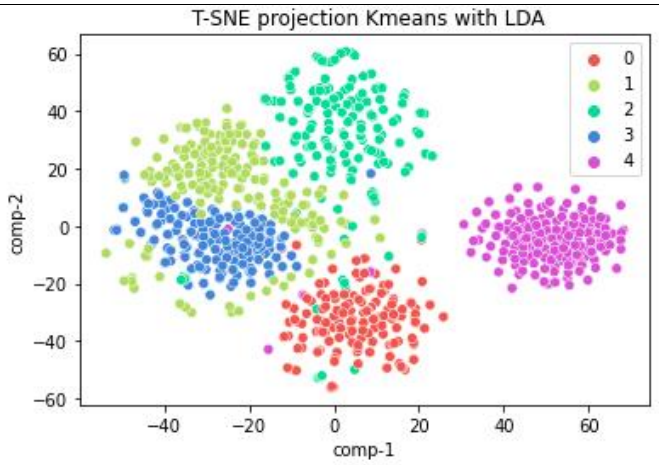
K-MEANS WITH LDA 69% V-Score



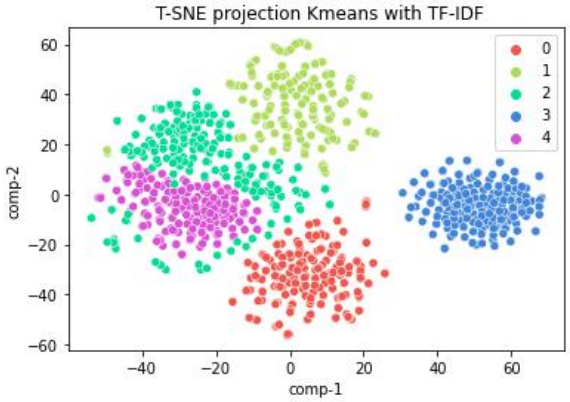
K-MEANS WITH TF-IDF 96% V-Score

k-means with BOW- true Y	K-means with BOW – Y hat
	
K-means with LDA- Y actual	**K-means with LDA- Y hat**

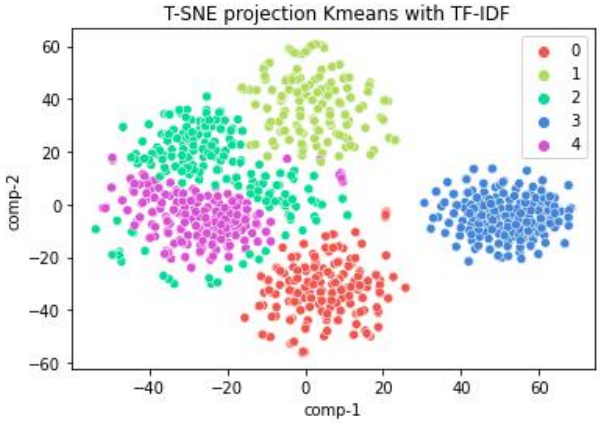




K-means with TF-IDF – Y actual

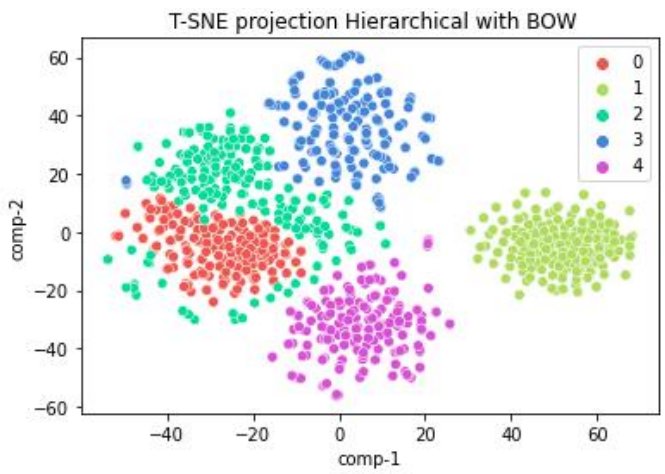


K-means with TF-IDF – Y hat

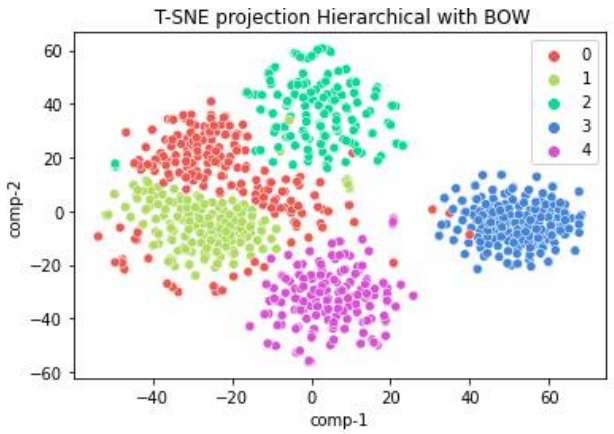


Hierarchical

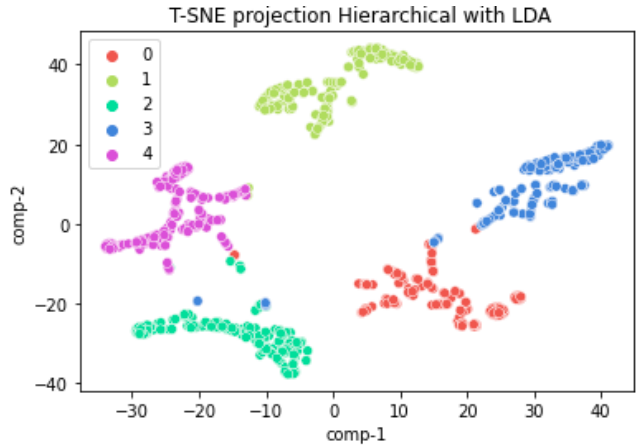
Hierarchical with BOW y-actual



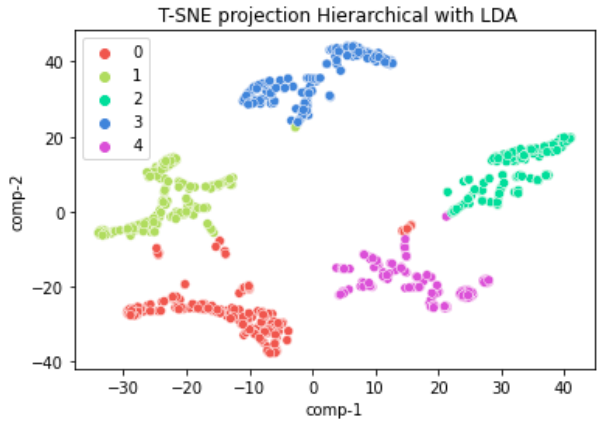
Hierarchical with BOW y-hat



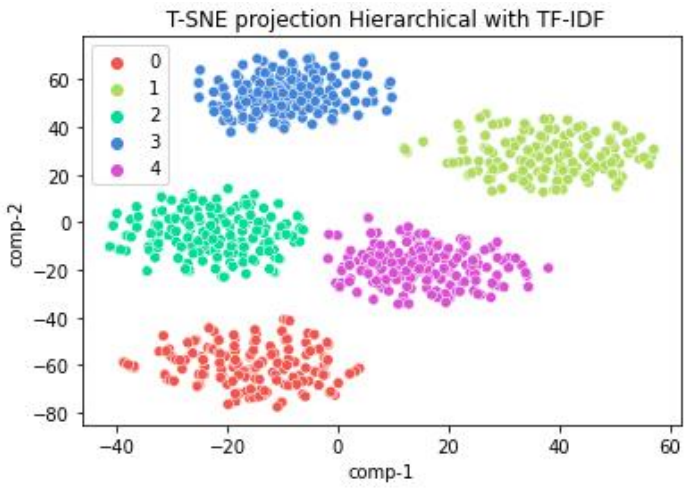
Hierarchical with LDA- Y-actual



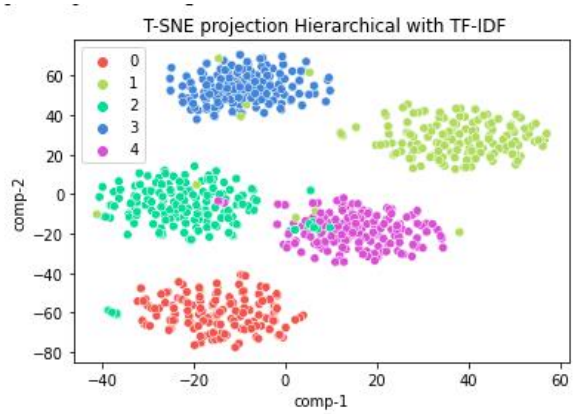
Hierarchical with LDA- Y-hat



Hierarchical with TF-IDF Y-actual

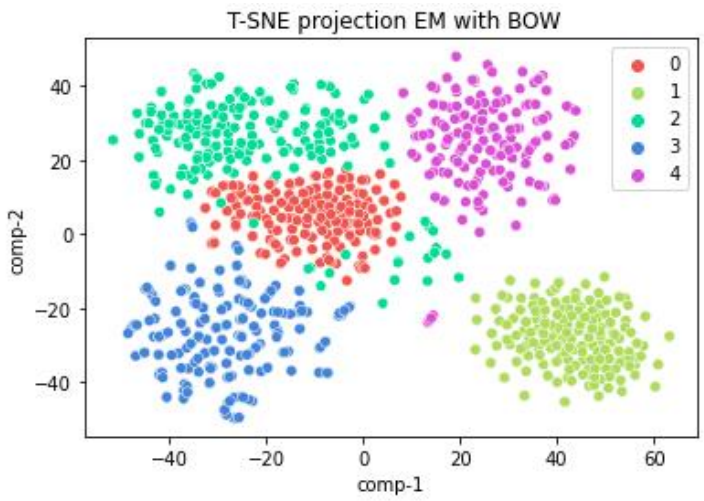


Hierarchical with TF-IDF Y-hat

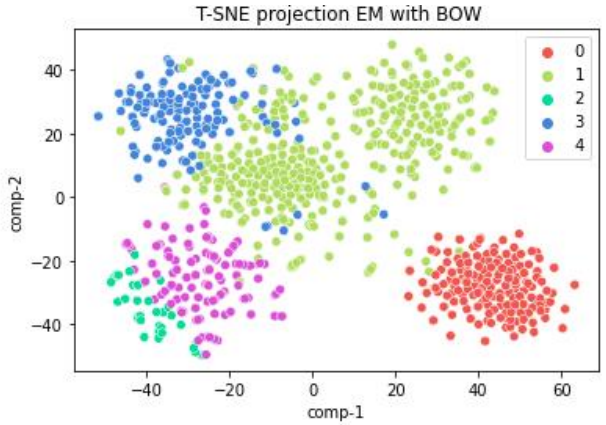


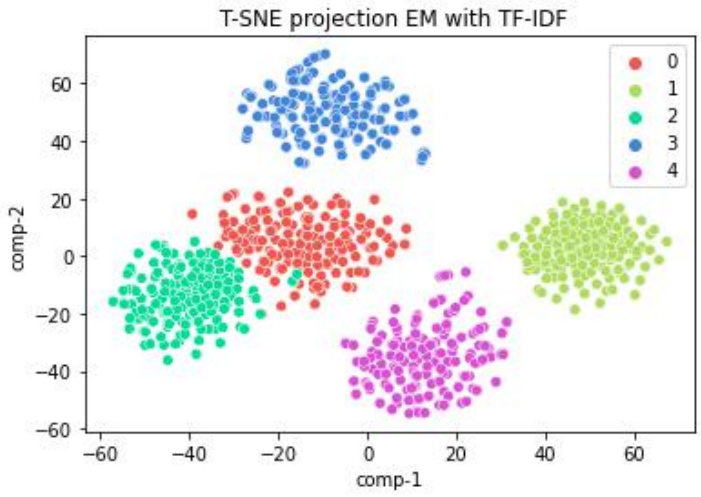
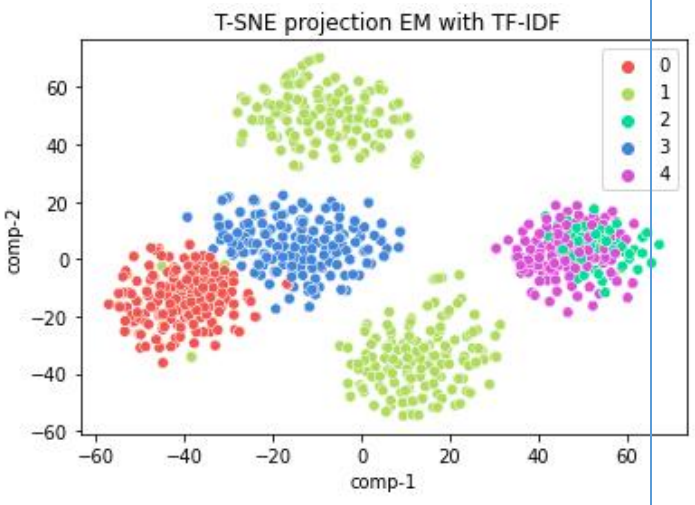
Gaussian matrix

EM with BOW (Y- actual)



EM with BOW (Y- hat)



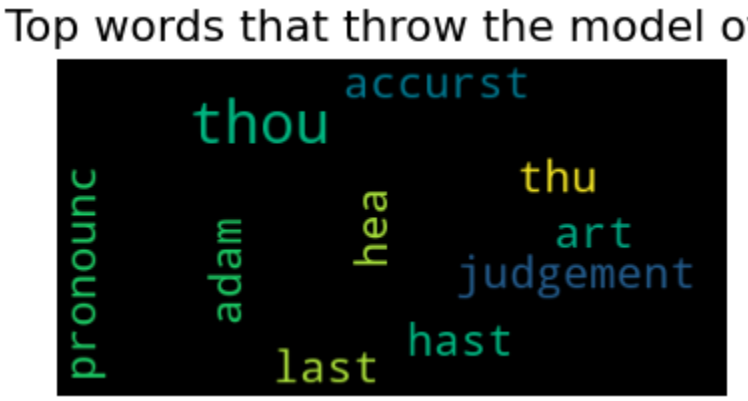
EM with TF-IDF (Y- actual)	EM with TF-IDF (Y- hat)
	

Perform Evaluations and Analysis of Bias and Variability:

The model	Silhouette	V-Score	Kappa Score
K-means with BOW	0.02	0.83	0.85
K-means with LDA	0.48	0.72	0.81
K-means with TF-IDF	0.03	0.99	0.99
Hierarchy with BOW	0.02	0.92	0.96
Hierarchy with LDA	0.45	0.78	0.84
Hierarchy with TF-IDF	0.03	0.99	0.99
EM with BOW and PCA	0.01	0.73	0.85
EM with TF-IDF and PCA	0.02	0.80	0.88
EM with LDA	0.27	0.58	0.63

Perform Error Analysis

These are the top ten words that occurred in all the clusters that had wrong classification



These are the top ten bigrams that occurred in all the clusters that had the wrong classification