Data:

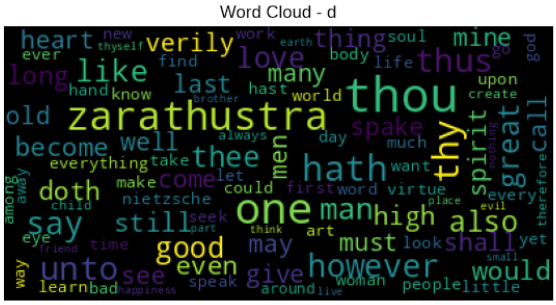
Our data consists of five books from the Gutenberg library. those books are:

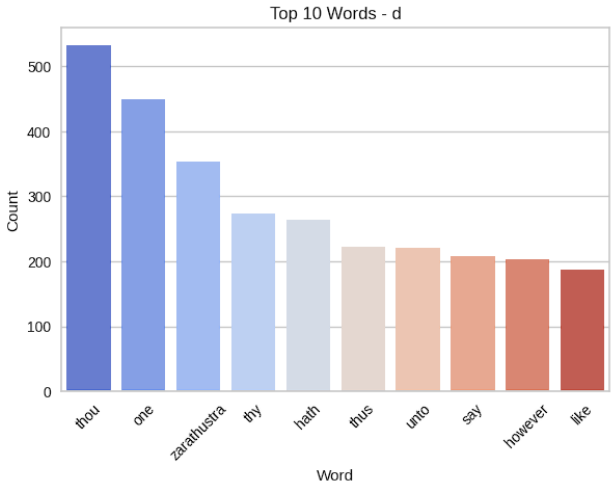
1. The Adventures of Pinocchio by Carlo Collodi
2. Les Misérables by Victor Hugo
3. Dracula by Bram Stoker
4. Thus Spoke Zarathustra by Friedrich Nietzsche
5. The Adventures of Sherlock Holmes by Arthur Conan Doyle

The books have Five different authors and are in Five different genres. After reading the books We start the data preparation by:

- tokenizing the book to get a list of separated words
- converting all words into lowercase words
- removing the stop words
- splitting each book to 200 samples, each sample containing 150 words
- lemmatizing every word then we combine all the samples into the complete data frame containing 1000 rows

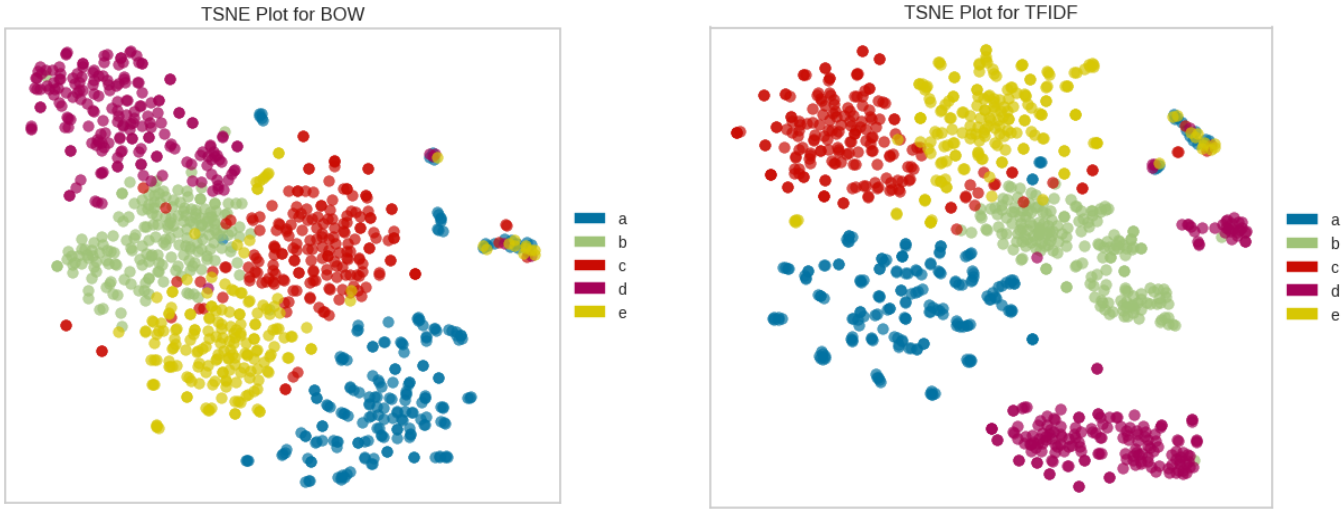
then we explore our data by plotting the most frequent 200 words of every book in a word cloud and then we plot the top 10 words to see how much they differ from other books.





Feature Engineering:

We used two different techniques, BOW and TFIDF, to transform our data and make it suitable for analysis. To improve the performance of our models, we also applied two dimensionality reduction methods (PCA and LDA), which helped us reduce the complexity of our data. Then we created TSNE plots, which allowed us to visualize the data in a way that highlighted clusters and allowed us to see the difference between clusters.



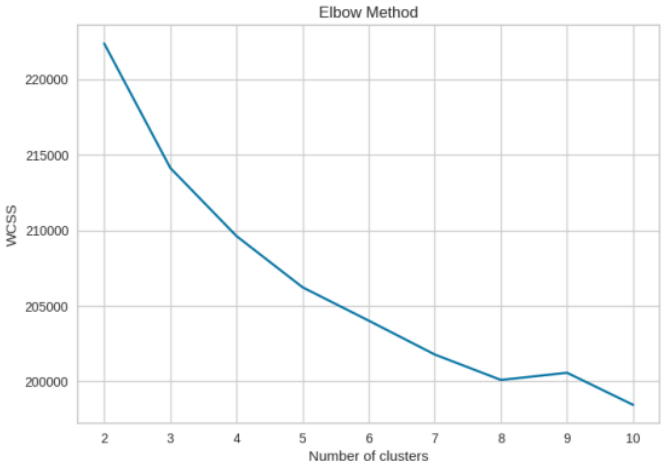
Modeling:

For each of our transformation methods (BOW, TFIDF, LDA), we run our models K-Means, Agglomerative Clustering, Gaussian Mixture and DB-Scan Models

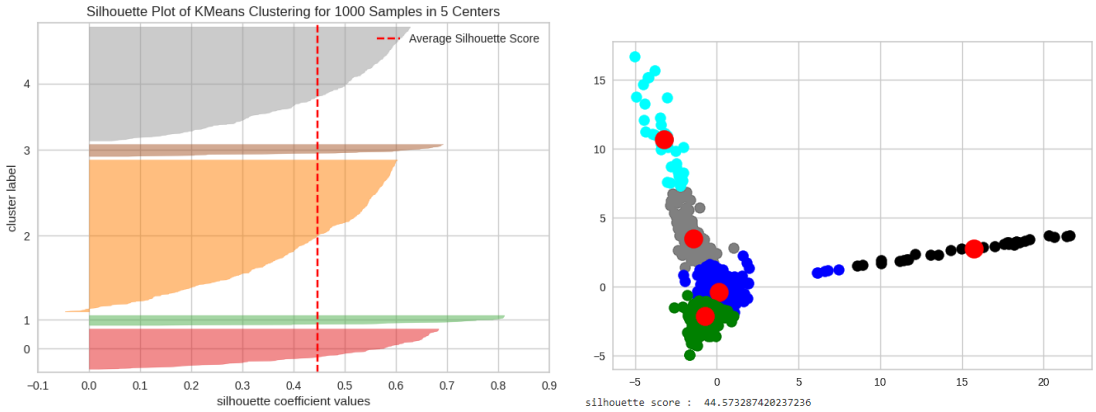
Model	Bag of Words	TFIDF	LDA
K-Means	silhouette: 44.57% Kappa: 0.24	silhouette: 63.77% Kappa: 0.3712	silhouette: 67.08% Kappa: 0.0237
AGG	silhouette: 61.16% Kappa: 0.0862	silhouette: 61.16% Kappa: 0.0862	silhouette: 61.16% Kappa: 0.0862
GMM	silhouette: 65.19% Kappa: 0.15125	silhouette: 65.19% Kappa: 0.15125	silhouette: 65.19% Kappa: 0.15125
DBSCAN	silhouette: 59.99% Kappa: 0.0401	silhouette: 59.99% Kappa: 0.0401	silhouette: 59.99% Kappa: 0.0401

- **K-Means:**

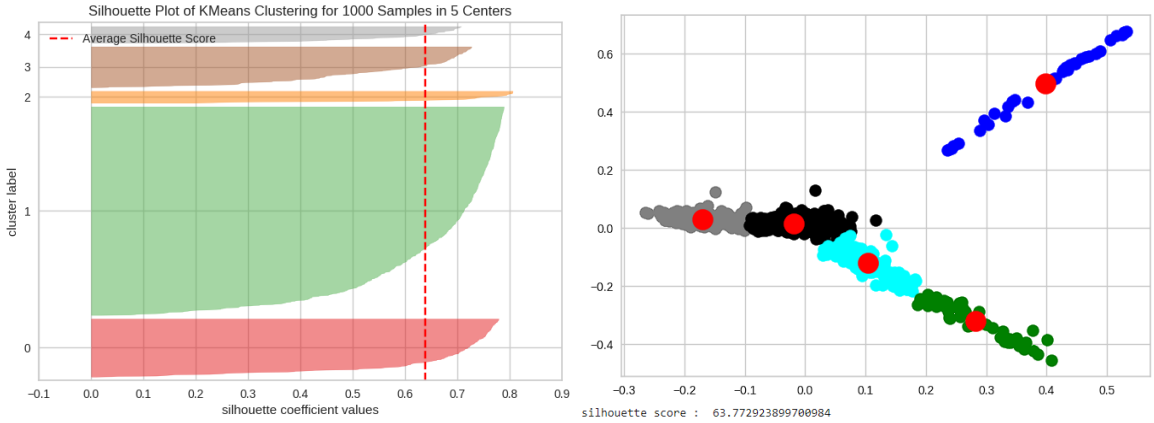
  First we run K-Means upon the bag of words representation with different numbers of K to decide the best K.



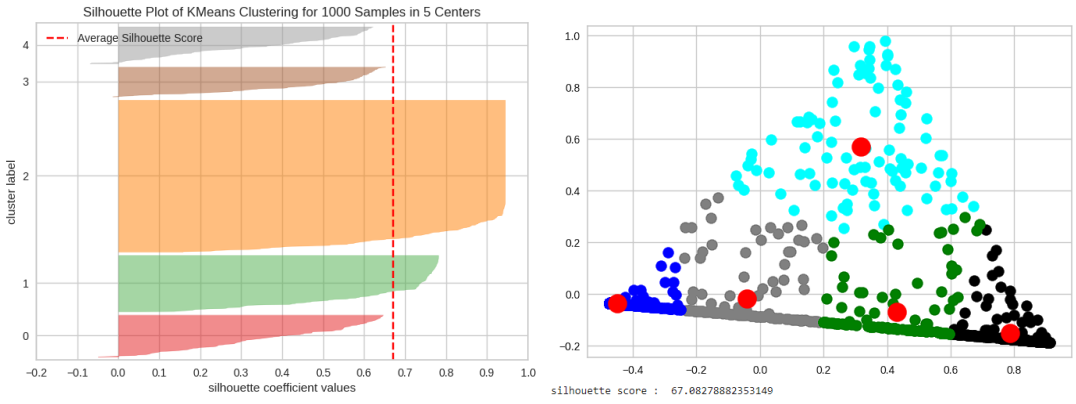
Then we run the K-Means with K equal 5 upon the bag of words representation and we print and visualize the silhouette score and visualize the scatter of clusters label predicted by the model



And for the TFIDF representation we do the same as above and we get a higher silhouette score



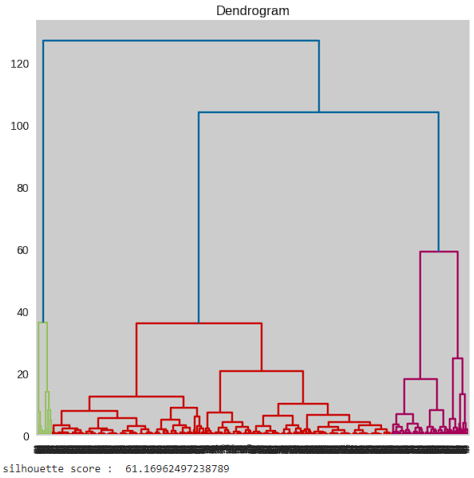
Finally, for the LDA representation we got almost the same silhouette score as the bag of words representation

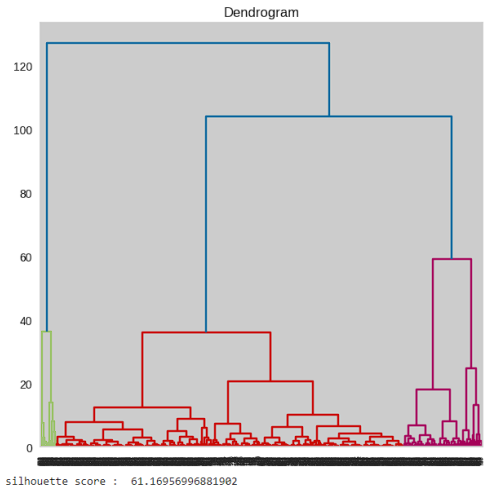


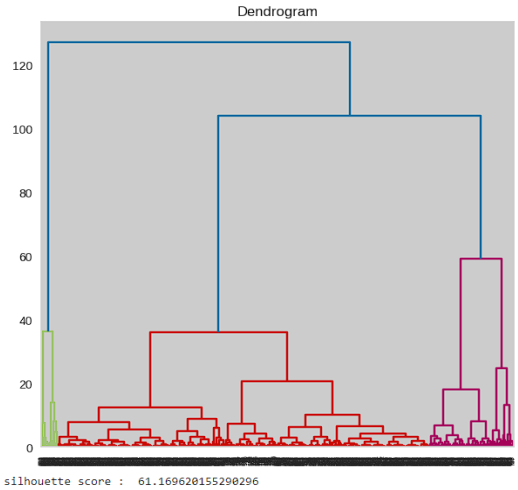
Comparing the 3 results to each other by the visualization we figure out that the clusters are NOT well separated specially in the bag of words and TFIDF representations but in the LDA we can see a clear separation between clusters

- **Agglomerative Clustering:**

  We run the Agglomerative clustering upon the 3 representation as before but here we found that the silhouette and kappa scores are almost the same and there is a slight change in it also we visualized the dendrogram of the 3 models and they are the same







- **Gaussian Mixture:**

  For the 3 representations we got the highest silhouette score between all models but coming to the kappa score we got the second worst score between all models

- **DBSCAN**:

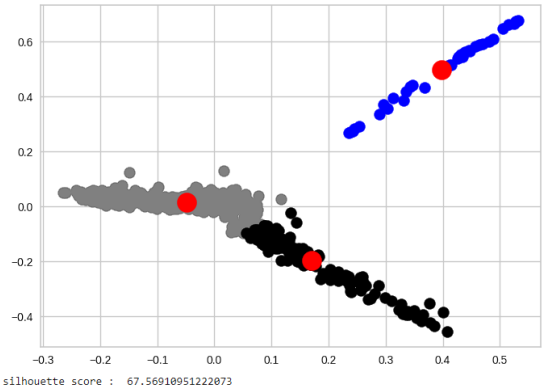
  We choose this model to automatically cluster our data without assigning any clusters number and we got a reasonable score for the silhouette and not much improvement in kappa than the Gaussian Mixture model

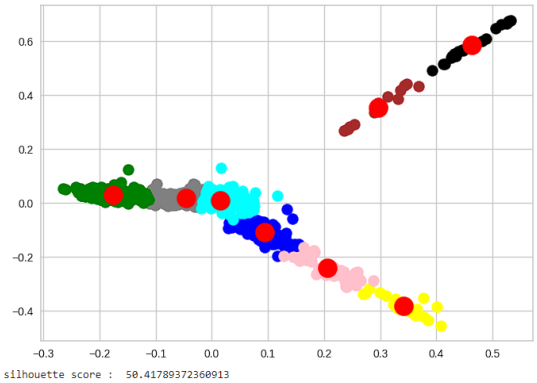
Champion Model:

We choose the K-Means model with the TFIDF transformation as it has best tradeoff between silhouette score and kappa score.

Error Analysis:

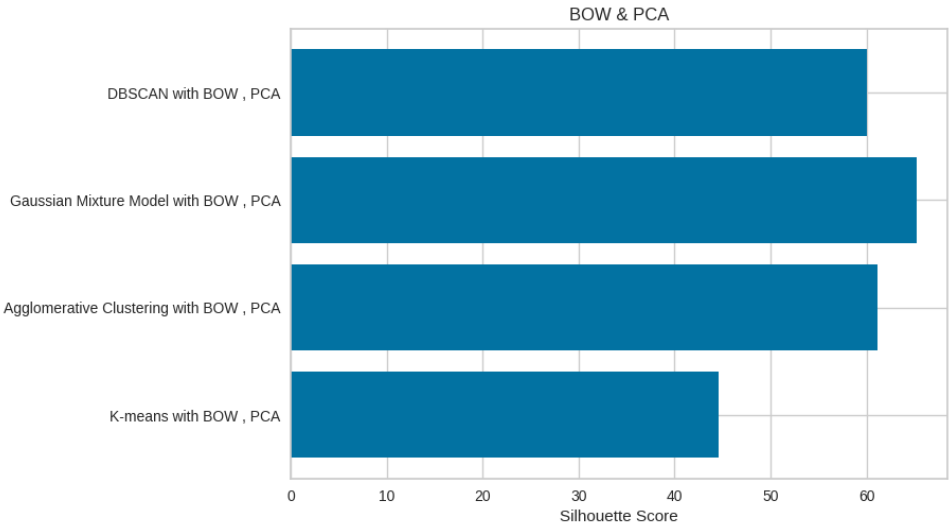
We changed the number of clusters of the champion model to K = 8 as determined by the elbow method seen before but it gave us a lower score than K = 5 then we tried K = 3 which gave us a higher score than K = 5 as expected.

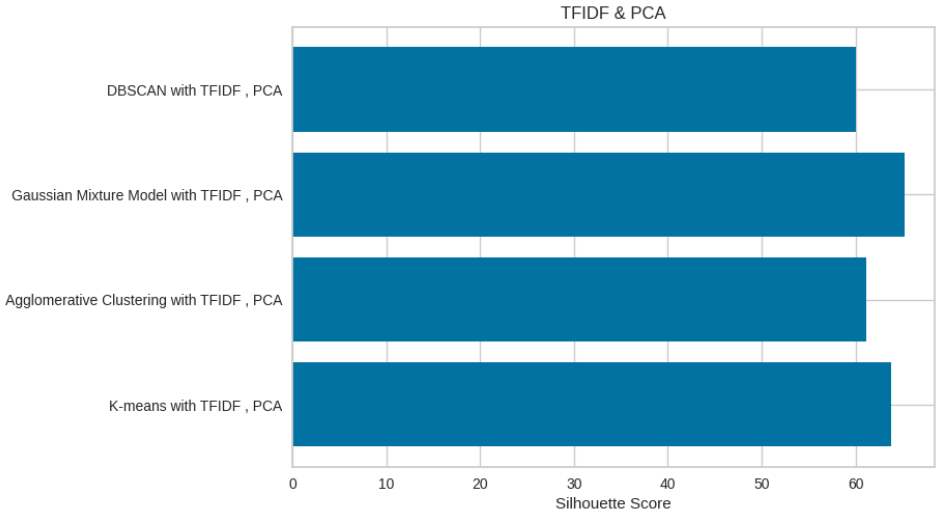


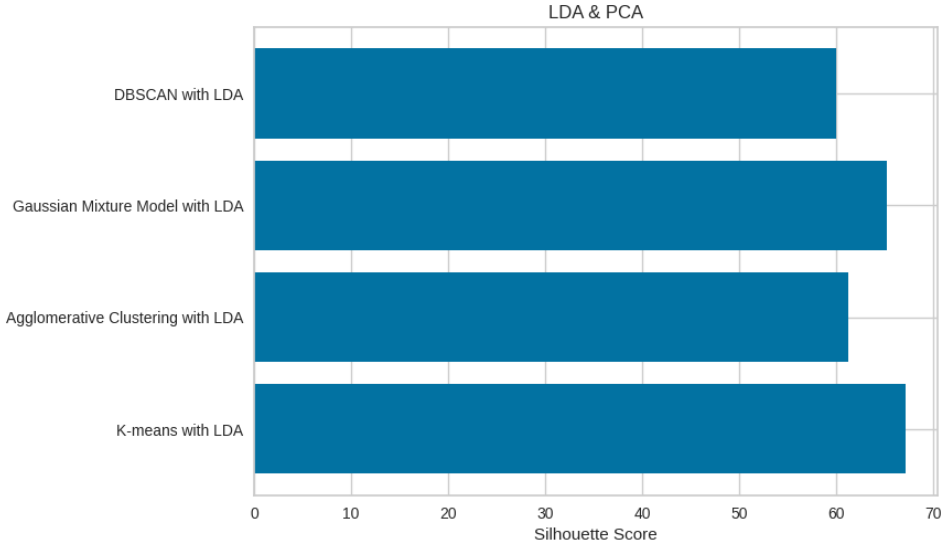


Testing Results:

For every model of the models we use, we output the Silhouette and Kappa scores and for each transformation technique, we plot the silhouette scores for all models in a bar plot







README:

There is ONE notebooks (book_clustering.ipynp) the notebook containsthe implementation of the clustering model with its own data importing, preprocessing and feature engineering.

the notebooks have commands to install and import the necessary libraries in order for the notebook to run correctly.