uOttawa

Report

Group Assignment 2

Data Science Applications

Group Number: G4

Team Members:

Abdallah Ragab, Hossam Mahmoud

Ahmed Ali Ziada, Shaimaa Mamdouh Ahmed

1. Overview

The main objective in this assignment is how to clustering five different books. We need to build machine learning models can perform this task on it. Our Goal is achieving this model with best performance by using different algorithms, packages, and tools of python.

2. Methodology

2.1 Import useful libraries

for reading and plotting such as pandas, NumPy, Sklearn, seaborn, matplotlib, etc.

2.2 Load 5 different books of different categories for different authors.

loading the Gutenberg dataset from NLTK package.

Selected books:

- Emma, Romance category
- Alice's Adventures in Wonderland, Fiction category
- The King James Version of the Bible, Religious category
- The Man Who Was Thursday, Mystery category
- Leaves of Grass, Drama category

2.3 Pre-processing of the Data

2.3.1 Cleaning

Using regular expression, the book title, chapter title, Volume title, empty lines, punctuation marks, and stopwords have been removed, then convert the book to small text.

2.3.2 Partitioning

creating random samples of 200 documents of each book and put them into a data frame, each record in this data frame contains 150 words for each document labeled with authors' names.

2.3.3 Label Encoding

Converting the names of the authors to numerical type such as 0, 1, 2, … etc.

2.4 Feature engineering

In this step several methods were preformed such as, The Bag of words (BOW) technique, Term Frequency-Inverse Document Frequency (TF-IDF), Term Frequency-Inverse Ngram, and Latent Dirichlet Allocation (LDA) to Convert text into vector.

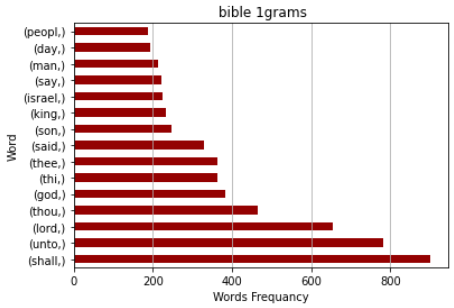
2.5 Subset the data

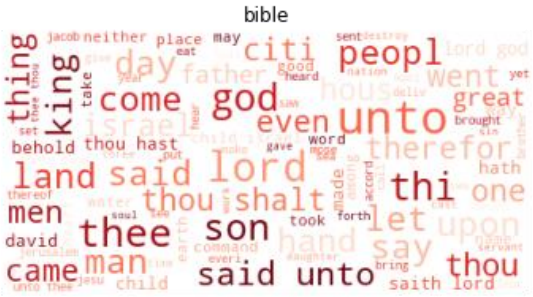
Define features which are words and target which is the book's title.

2.6 Exploring the data

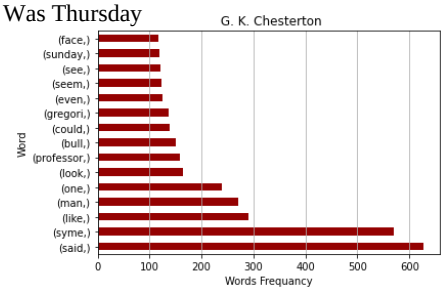
Plotting the most repeating words in each book, and show the word cloud for each book

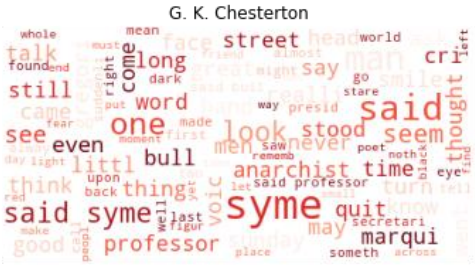
The King James Version of the Bible



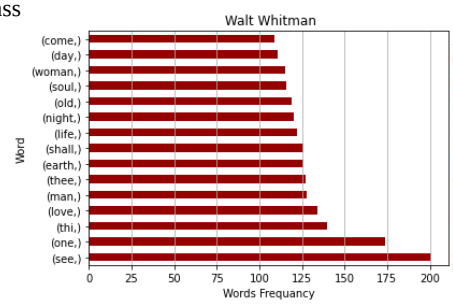


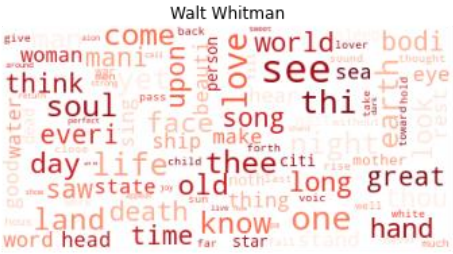
The Man Who Was Thursday



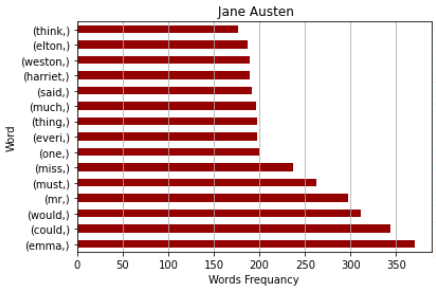


Leaves of Grass

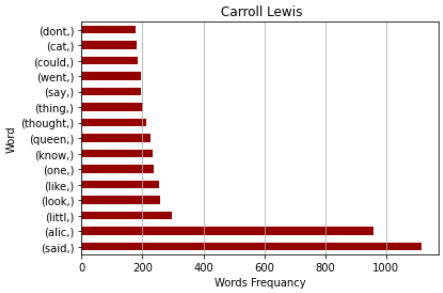


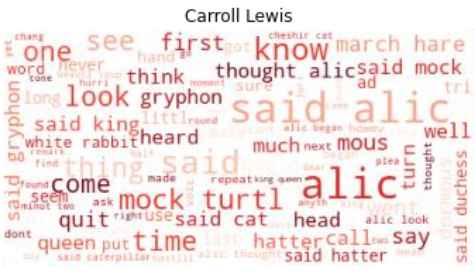


Alice's Adventures in Wonderland



Jane Austen





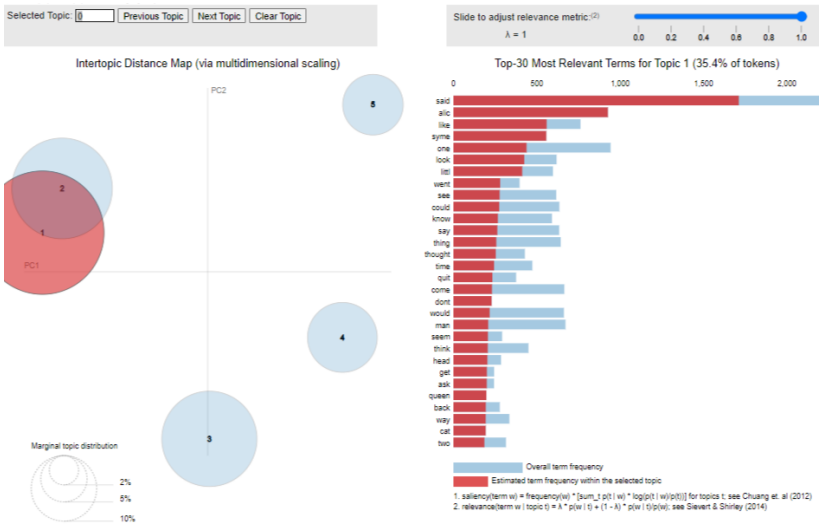
2.7 Models

Multiple models were preformed:

- Kmean
- EM Model
- Hierarchical

These models trained by four different methods of Text transformation

- BOW
- TF-IDF
- N-Gram
- PCA
- Word2Vec
- **LDA**



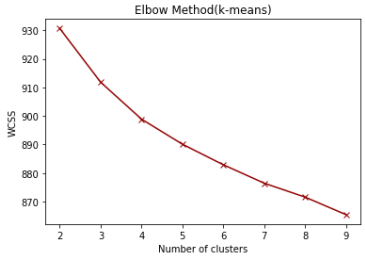
2.7.1 Kmean model

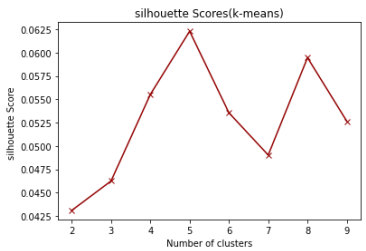
Using elbow method and silhouette score as shown below, they have shown that best K is 5, as we knew from the dataset.

- Train kmean model with k=5.
- Mapping between books and the output of the model by assigning the cluster to most frequent book's partitions in this cluster.
- Calculate kappa and silhouette and other metrics.
- Visualize data with wrong points.

2.7.1.1 Using TF-IDF data.

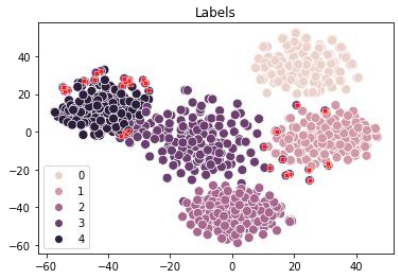
2.7.1.1.1 Plot elbow method and silhouette score





2.7.1.1.2 Metrics scores and Visualize data with wrong points

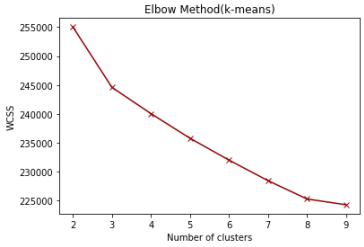
Kappa	0.948
Silhouette Coefficient	0.062
Homogeneity	0.907
Completeness	0.910
V-measure	0.908
Adjusted Rand-Index	0.899

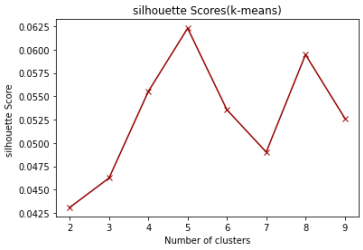


- In this model cluster our data in suitable way and TF_IDF make data easier to represent in multiple clusters.

2.7.1.2 Using BOW data

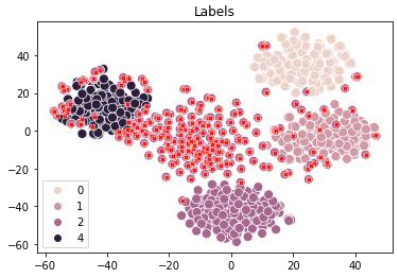
2.7.1.2.1 Plot elbow method and silhouette score





2.7.1.2.2 Metrics scores and Visualize data with wrong points

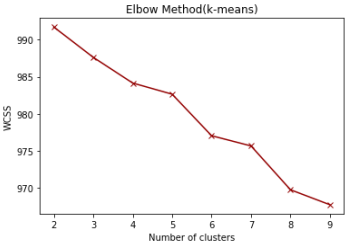
Kappa	0.641
Silhouette Coefficient	0.074
Homogeneity	0.615
Completeness	0.0.783
V-measure	0.689
Adjusted Rand-Index	0.530

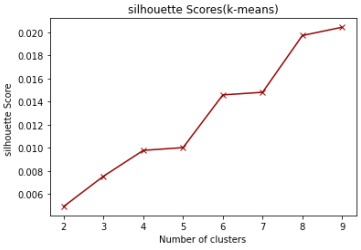


- In this model it treats 2 books as 1 cluster and this drives to low Kappa score.

2.7.1.3 Using 2Gram data.

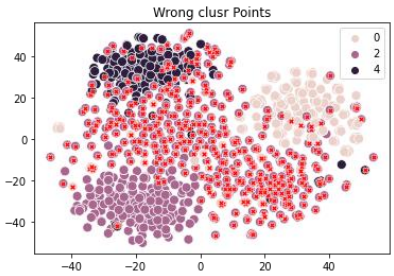
2.7.1.3.1 Plot elbow method and silhouette score





2.7.1.3.2 Metrics scores and Visualize data with wrong points

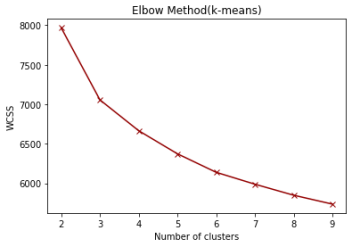
Kappa	0.382
Silhouette Coefficient	0.006
Homogeneity	0.314
Completeness	0.551
V-measure	0.400
Adjusted Rand-Index	0.258

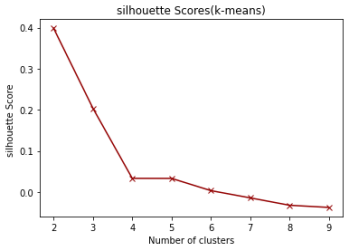


- In this model it treats 3 books as 1 cluster and this drives to low Kappa score

2.7.1.4 Using Word2Vec data

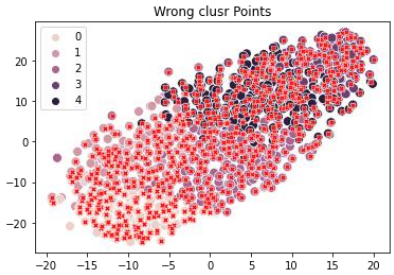
2.7.1.4.1 Plot elbow method and silhouette score





2.7.1.4.2 Metrics scores and Visualize data with wrong points

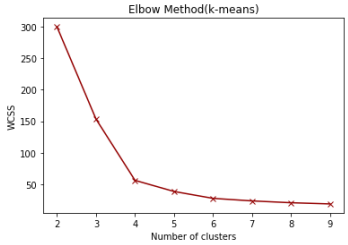
Kappa	0.040
Silhouette Coefficient	0.033
Homogeneity	0.004
Completeness	0.004
V-measure	0.004
Adjusted Rand-Index	-0.001

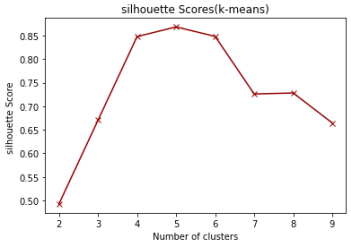


- Kmeans performed parley with Word2vec it got confused between books.

2.7.1.5 Using LDA data

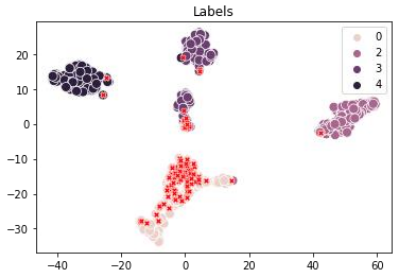
2.7.1.5.1 Plot elbow method and silhouette score





2.7.1.5.2 Metrics scores and Visualize data with wrong points

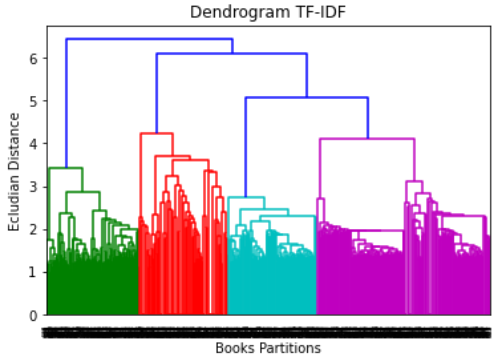
Kappa	0.741
Silhouette Coefficient	0.832
Homogeneity	0.760
Completeness	0.906
V-measure	0.827
Adjusted Rand-Index	0.739



2.7.2 Hierarchical Model

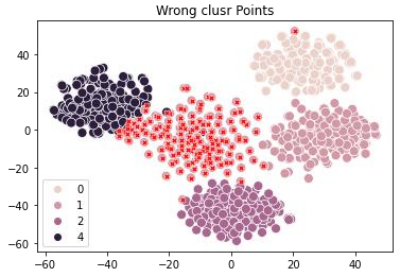
2.7.2.1 Using TF-IDF data

2.7.2.1.1 Plot the dendrogram.



2.7.2.1.2 Metrics scores and Visualize data with wrong points.

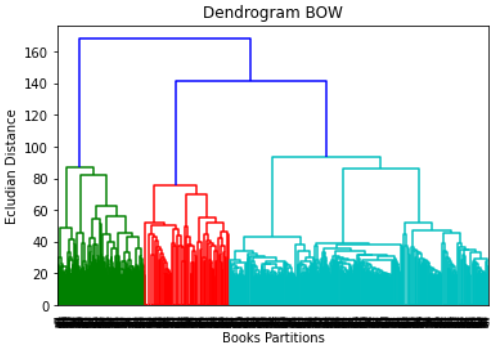
Kappa	0.746
Silhouette Coefficient	0.056
Homogeneity	0.789
Completeness	0.949
V-measure	0.862
Adjusted Rand-Index	0.759



- In this model it treats 2 books as 1 cluster and this drives to low Kappa score

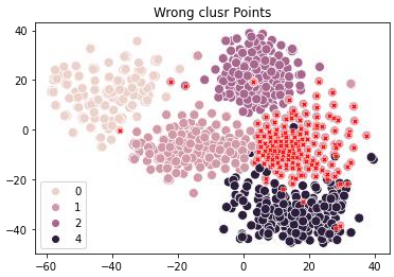
2.7.2.2 Using BOW data

2.7.2.2.1 Plot the dendrogram.



2.7.2.2.2 Metrics scores and Visualize data with wrong points.

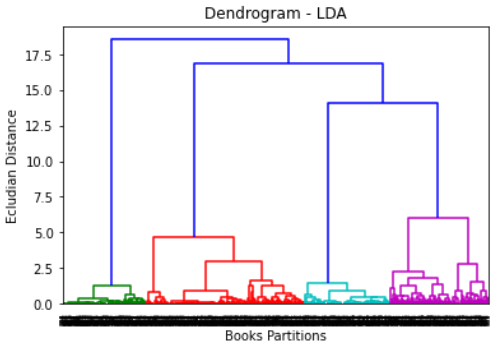
Kappa	0.731
Silhouette Coefficient	0.083
Homogeneity	0.757
Completeness	0.949
V-measure	0.829
Adjusted Rand-Index	0.728



- In this model it treats 2 books as 1 cluster and this drives to low Kappa score

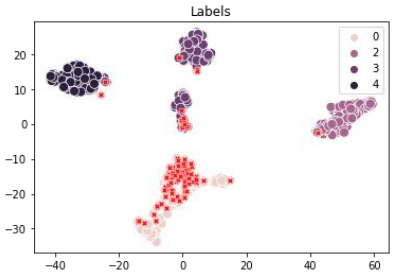
2.7.2.3 Using LDA data

2.7.2.3.1 Plot the dendrogram.



2.7.2.3.2 Metrics scores and Visualize data with wrong points.

Kappa	0.740
Silhouette Coefficient	0.810
Homogeneity	0.763
Completeness	0.911
V-measure	0.830
Adjusted Rand-Index	0.736



- In this model it treats 2 books as 1 cluster and this drives to low Kappa score

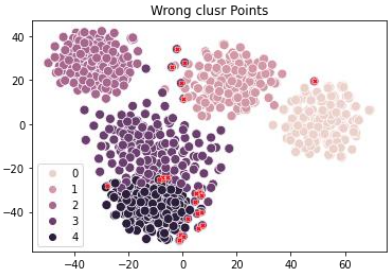
2.7.3 EM Model

First, we will perform PCA on Dataset to avoid memory leak problem with PCA component = 600 which represent 97 % of our dataset variance. Then perform EM on this data

2.7.3.1 Using TF-IDF data

2.7.3.1.1 Metrics scores and Visualize data with wrong points.

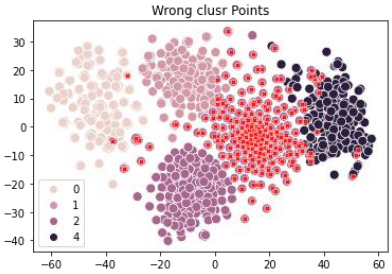
Kappa	0.740
Silhouette Coefficient	0.057
Homogeneity	0.783
Completeness	0.945
V-measure	0.856
Adjusted Rand-Index	0.746



2.7.3.2 Using BOW data

2.7.3.2.1 Metrics scores and Visualize data with wrong points.

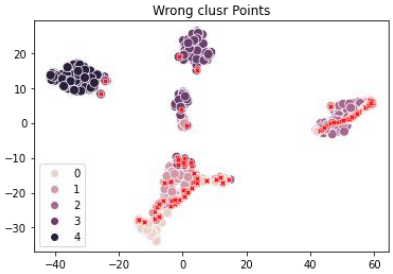
Kappa	0.640
Silhouette Coefficient	0.075
Homogeneity	0.612
Completeness	0.780
V-measure	0.686
Adjusted Rand-Index	0.527



2.7.3.3 Using LDA data

2.7.3.3.1 Metrics scores and Visualize data with wrong points.

Kappa	0.791
Silhouette Coefficient	0.377
Homogeneity	0.732
Completeness	0.742
V-measure	0.737
Adjusted Rand-Index	0.669



2.8 Compare models

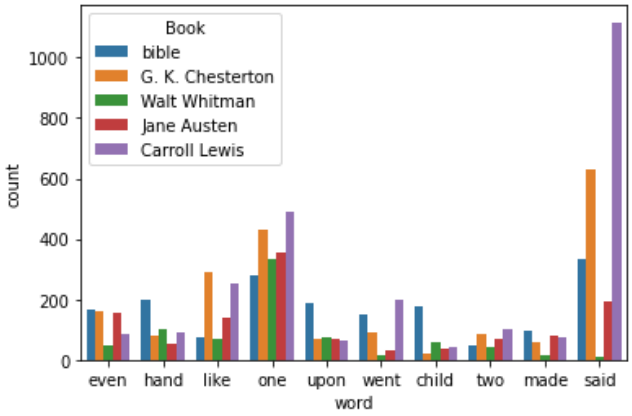
Models	Kappa	Silhouette	V-measure
KMeans_ TFIDF	0.948	0.062	0.908
KMeans_ BOW	0.641	0.074	0.689
KMean_2Gram	0.382	0.006	0.400
KMean_word2vec	0.040	0.033	0.004
KMean_LDA	0.741	0.832	0.827
Hierarchical_ TFIDF	0.746	0.056	0.862
Hierarchical_ BOW	0.731	0.083	0.829
Hierarchical_ LDA	0.740	0.810	0.830
EM_ PCA_TFIDF	0.740	0.057	0.856
EM_PCA_BOW	0.640	0.075	0.686
EM_LDA	0.791	0.377	0.737

2.9 Champion Model

By comparing between all accuracies for each model KMeans **Model** that trained on TF_IDF had the highest Kappa score and the nearest model to true labels, this model is our champion model.

3. Error Analysis

Get number of the wrong points which is **42 points** and invers the transform into words then get most **10 frequent** words that model put it in wrong clusters and how frequent these words in all books.



4. Conclusion

We have learned many new things during this assignment, we have deep dive in feature engineering techniques by applying new techniques like Word2Vec and LDA.

And we have learnt many different clustering techniques like K-means, Expectation-Maximization (EM) and hierarchal clustering with each of vectorization method.

And we have applied different evaluation techniques like kappa, silhouette, and V-measure to decide which model will be the champion model.

After that we have analysis our champion model results to determine which partitions and which words in these partitions that threw the machine off.