ASSIGNMENT 2

Group 4

ABSTRACT

The overall objective of this assignment was to produce similar text clusters and compare them; analyze pros and cons of algorithms and generate and communicate the insights.

AUTHORS:

Ifiok Charles (0300041771)

Kalada Reagan Tuwotamuno (8489637)

Stevy Kuimi (8845324)

Gayatri (300095631)

SUPERVISOR:

Prof. Arya Rahgozar

GNG 5125 X

Contents

Introduction **3**
Data preparation **4**
Data **4**
Data preparation, preprocessing and cleansing **4**
Feature Engineering **5**
Modelling **7**
Models: **7**
Performance and Evaluation: **8**
Error Analysis **10**
Visualisatuon ..... Error! Bookmark not defined.

Table of Tables

Table 1 Advantages and disadvantages of the features used 6
Table 2 Kappa Value interpretation 8
Table 3 Kappa and Silhouette values for all models 9
Table 4 Adjusted Rand Scores 10

Table of figures

Figure 1 Graphical representation of the results. 9
Figure 2 Cluster plot for BOW 9
Figure 3 Cluster plot for TF-IDF 9

Introduction

The overall objective of this assignment was to produce clustering predictions and compare them; analyze pros and cons of algorithms and generate and communicate the insights.

Data preparation

Data

The task of this assignment is to classify segments of books by authors. This was done by choosing five random books from different genre from the Gutenberg online libraries. They were chosen in the categories of mythology, music, engineering, physics and poetry. The table below shows the names of the authors of each book chosen.

Genre	**Authors**
Mythology	Sir James George Frazer
Music	Magaret Blake-Alverson
Engineering	Charles M. Hortons
Physics	Amos Emerson Dolbear
Poetry	John Milton

Data preparation, preprocessing and cleansing

These are the steps we followed in writing our code:

- Upload the data to the data frame,
- Tokenize the data to make sure the characters are separated,
- Remove stop words to lower the dimensional space and help with the collocation
- Stemming: Basically, this convert words into their root form by reducing the difference between their inflected forms,
- create random samples of 200 documents of each book, each sample with a 150 words.
- Create a corpus, meaning putting all the document samples in the same data frame and proceed further cleaning, which involves only keeping letters of the alphabet from a-z.
- Removing all numbers, white spaces and punctuations, and putting all the words into lower cases. This was necessary for the predictor to avoid confusion.

Feature Engineering

- **BOW:** Bags of Words. This is a representation of text that describes the occurrence of words within a document. This is the most popular technique used to convert categorical features to numerical ones.
- **TF-IDF:** Term Frequency-Inverse Document Frequency: this is another way or technique used for occurrence of words. TF-IDF measures relevance, not frequency. The TF part divides the number of occurrences of each word in the document by the total number of words in the document, while the IDF does the downscaling of weight for words that occur in many documents in the corpus. For example, if the words like 'the', 'and' appears in all documents, those will be systematically discounted. Each word's TF-IDF relevance is a normalized data format that also adds up to one.
- **LDA:** Latent Dirichlet Allocation. It is a form of unsupervised learning that views documents as bags of words (ie order does not matter). (Blei, et al., 2003) define it as a generative probabilistic model of a corpus with the idea of representing each document as a random mixture over latent topics, each topic being characterize by a distribution over words.

	BAG OF WORDS	**TF-IDF**	**LDA**
Advantages	Easy to understand Easy to implement	• suitable in comparing two documents suitable for long • document	• Can be embedded in more complicated models • Data-generating distribution can be changed
Dis-advantages	• Not suitable for long documents • Do not consider the semantic relation between words • Curse of dimensionality	• Less informative for assessing occurrence in long document • The dependence on BOW is a liability	• Unsupervised learning makes it difficult to evaluate the overall quality of a model. • Not suitable for short documents • The number of topics must be fixed and known

Table 1 Advantages and disadvantages of the features used

Modelling

Models:

Clustering is the process of grouping data according to their similarities. Three models were used to perform the task: K-means, Agglomerative-cluster and Gaussian Mixture Model with Expectation Maximization.

- **K-means.** This is the most used form of unsupervised machine learning technique for clustering. It has proven to be very fast and simple. K represent the number of clusters and these are the steps of the algorithm:
    1. Select the number of clusters, K
    2. Take each point and find the nearest centroid
    3. Match each point to the closest centroid again
    4. Repeat until the clusters cannot be improved anymore.
- **A-cluster:**. The main difference between K-means and A-cluster is that A-cluster do not initialize with random centroids.
    1. Take each point of the dataset as cluster
    2. Search and combine two closest points in one cluster
    3. Repeat until there is remaining only one big cluster
- **Gaussian Mixture Models with EM: Expectation-Maximization ,** Gaussian mixture gives more flexibility than K-means, the assumption made here are that the data point are gaussian distributed. The parameters of the gaussian are found using the Expectation-Maximization, which is an optimization algorithm. The algorithm goes as follow:
    1. Select the number of clusters
    2. Randomly initialize the gaussian distribution parameters for each cluster
    3. Compute the probability of each data point belonging to a cluster,
    4. Compute new set of parameters for the Gaussian distributions in a way to maximize the probabilities of data points within the cluster. This is done by using the weighted sum of the data points positions, where the weights are the probabilities of the data point belonging in that particular cluster,
    5. Repeat step 3 and 4 until convergence.

Performance and Evaluation:

For the evaluation, we used two techniques:

- **Cohen's Kappa:** it ranges from -1 to 1. Table 2 below shows how to interpret the kappa value.

Kappa Value	Agreement
< 0	Less than chance agreement
0.01-0.20	Slight Agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1	Almost perfect agreement

Table 2 Kappa Value interpretation

- **Silhouette:** Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]. Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

  This being said, Table 3 below shows the resulting values of Kappa and Silhouette for models with each feature engineering. Based on the interpretation theory above, it can be observed that K-means has a very good performance with LDA but did not perform that well with BOW. So our champion is K-means with LDA and the challenger is the EM with LDA.

	Kappa			Silhouette		
	BOW	**TF-IDF**	**LDA**	**BOW**	**TF-IDF**	**LDA**
K-means	-0.25	1.0	**1.0**	0.0486	0.0347	**0.7847**
A-Cluster	-0.25	0.24875	-0.00375	0.0486	0.0346	0.7821
EM	0.39	0.0	0.26875	0.0381	0.0347	0.7386

Table 3 Kappa and Silhouette values for all models

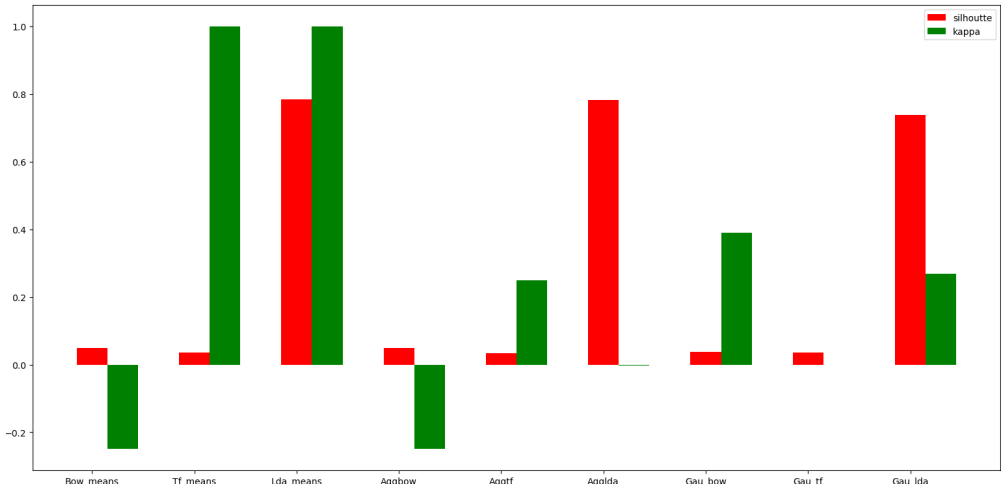


Figure 1 Graphical representation of the results.

Figure 2 Cluster plot for BOW

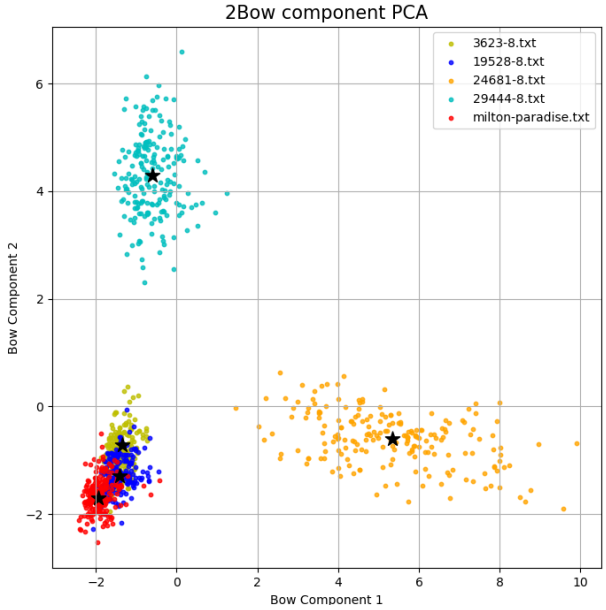


Figure 3 Cluster plot for TF-IDF

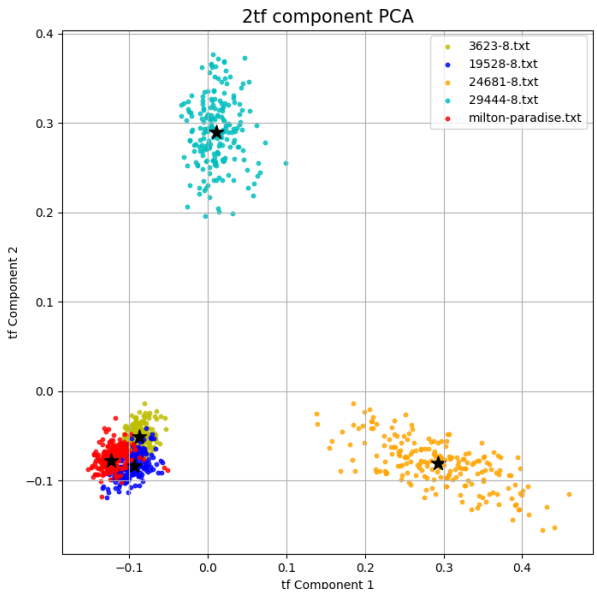


Figure 1 show the cluster obtained with BOW and K-means. It can be observed that the yellow, blue and red clusters are overlapping, this is an indication that they have similar words between them. The same is observed on Figure 2 below which represents the clusters plot for K-means with TF-IDF.

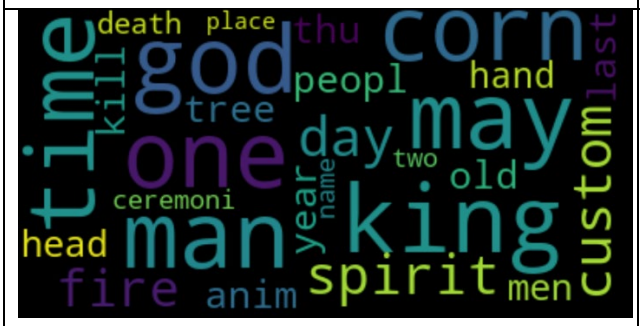
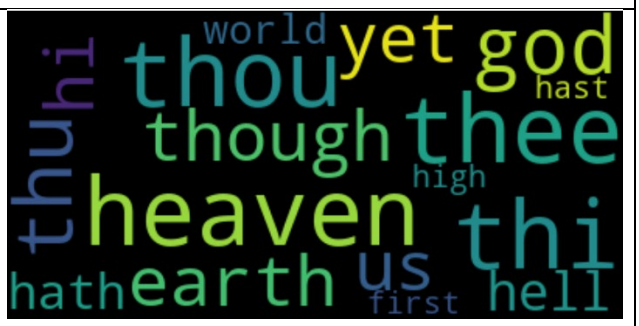
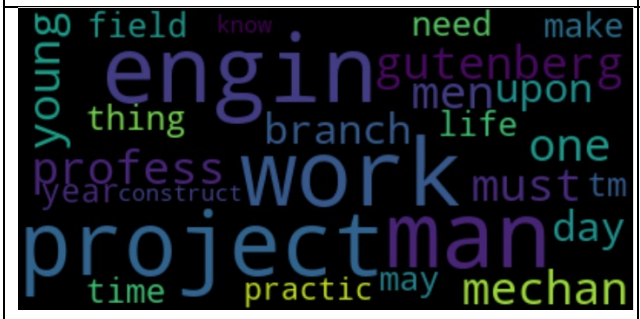
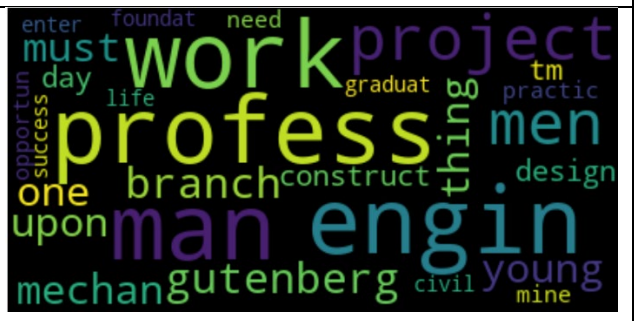
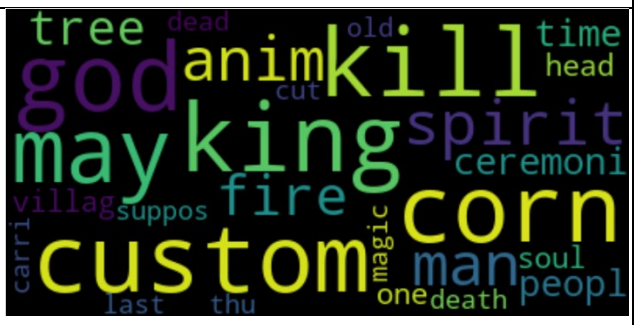
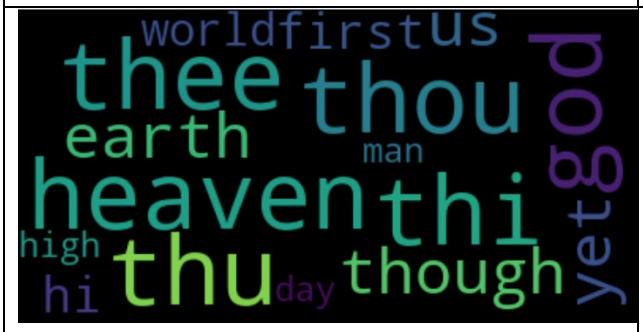
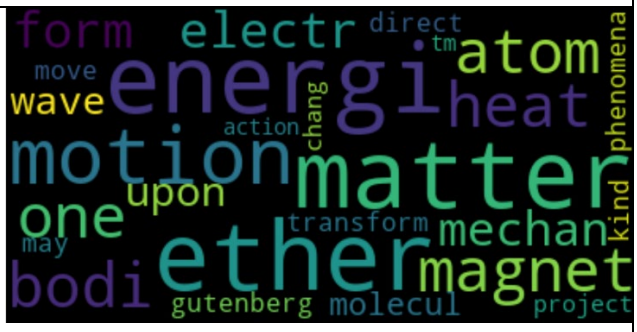
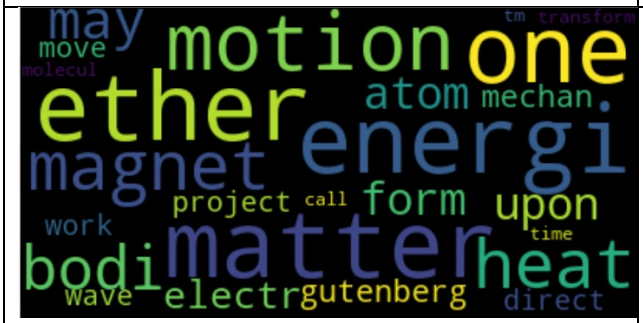
Adjusted Rand Scores

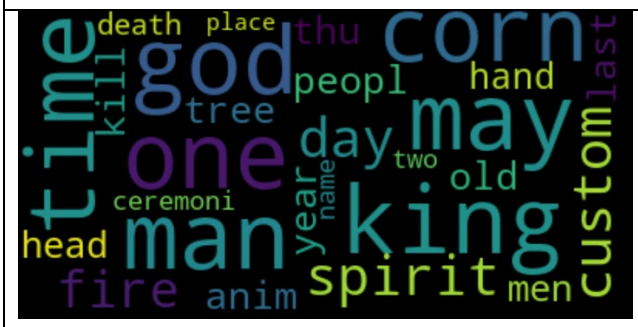
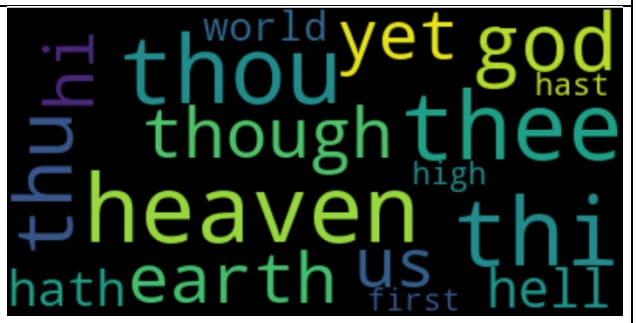
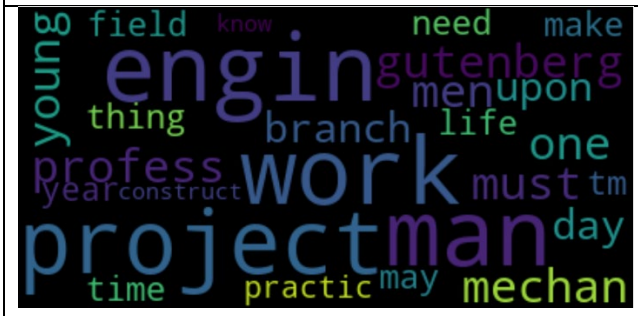
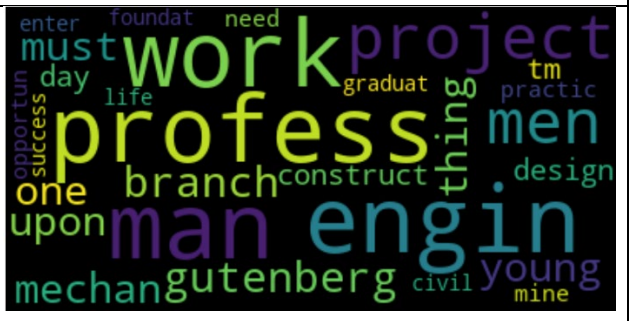
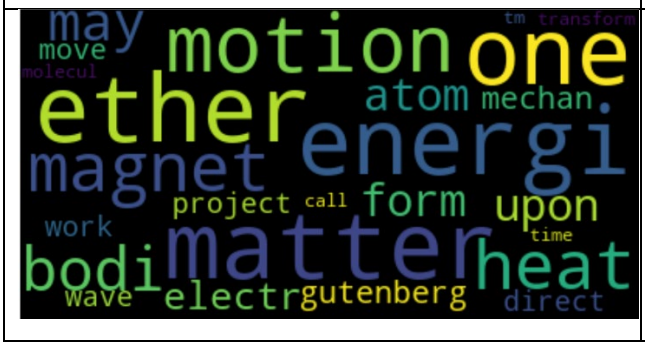
	BOW	TF-IDF	LDA
K-means	1	1	1
GMM/EM	0.715	1	0.951
A-cluster	1	0.997	0.990

Table 4 Adjusted Rand Scores

Error Analysis

For the error analysis, we looked for the most occurring words in each cluster and compared them for similarity. From the cluster plot above, there are similarities between clusters 0, 1 and 4. From the BOW word cloud table below, we can indeed see similar occurring words between cloud 0,1 and 4. This is one of the reasons why the K-means BOW model is very inaccurate.

BOW wordcloud by cluster from 0 to 5	TF-IDF wordcloud by cluster from 0 to 5
	
	
	
	
	

BOW wordcloud by cluster 0,1 and 4	TF-IDF wordcloud by cluster 0,1 and 4
	
	
	

References

Blei, D. M., Ng, A. Y. & Jordan, M. I., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3,* Volume 3, pp. 993-1022.