CLUSTERING

Assignment 2

uOttawa

Faculté de génie
Faculty of Engineering

Ahmed Hallaba **Bassel Hamshary**
Kamel El-Sehly

JUNE 19, 2021
UNIVERSITY OF OTTAWA
Ottawa, Canada

Table of Contents

Table of Figures

1. Objectives

The overall aim is to produce similar clusters and compare them; analyze the pros and cons of algorithms, generate, and communicate the insights.

2. Requirements

Take five different samples of Gutenberg digital books, which are all of five different genres and of five different authors, that are semantically similar. Separate and set aside unbiased random partitions for clustering.

3. Procedures

3.1. Data Preparation

The task of this assignment is to cluster segments of books by authors. This was done by choosing five random books from different genre from the Gutenberg corpus.

Genre	Book Name	Authors
Mythology	The Golden Bough	Sir James George Frazer
Music	Sixty Years of California Song	Magaret Blake-Alverson
Engineering	Opportunities in Engineering	Charles M. Hortons
Physics	The Machinery of the Universe	Amos Emerson Dolbear
Poetry	Paradise Lost	John Milton

3.1.1. Data Preprocessing

Steps of data preprocessing in the 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 converts 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 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.

3.1.2. 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. 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 documents.	• 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.

3.2. Clustering 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.

3.2.1. 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.

3.2.2. 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

3.2.3. Gaussian Mixture Models with Expectation-Maximization

Gaussian mixture gives more flexibility than K-means, the assumptions 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.

3.3. Performance Evaluation

Cohen's Kappa: Cohen's kappa is a metric often used to assess the agreement between two raters. we could use Cohen's kappa to compare the machine learning model predictions with the manually established ratings (real data). It ranges from -1 to 1.

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

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.

Table 1: Kappa and Silhouette values for all models

	Kappa			Silhouette		
	BOW	**TF-IDF**	**LDA**	**BOW**	**TF-IDF**	**LDA**
K-means	-0.213	0.235	0.041	0.057	0.059	0.471
A-Cluster	0.515	0.25	-0.053	0.086	0.06	0.465
EM	0.061	-0.25	-0.047	0.057	0.059	0.246

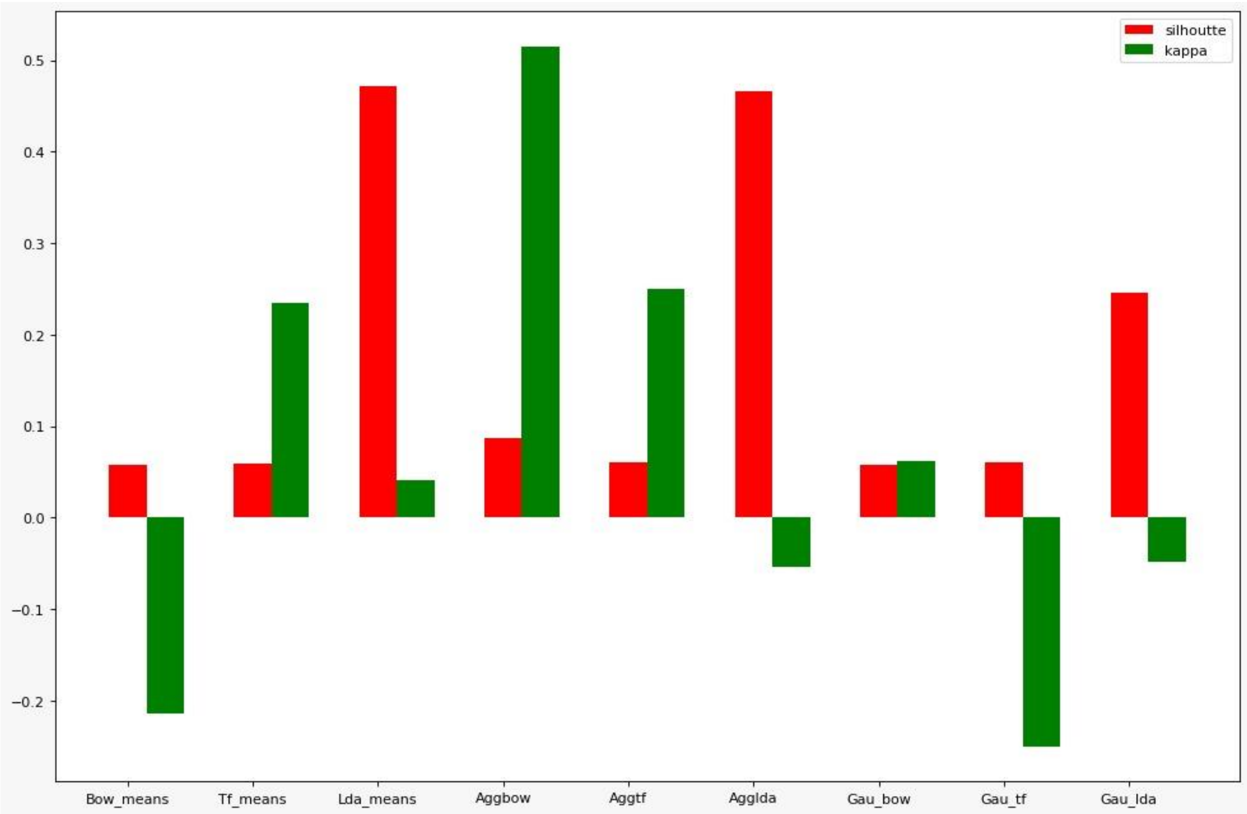


Figure 1: Graphical representation of Silhouette and Kappa results

Adjusted Rand Scores is a function that measures the similarity of the two assignments.

	BOW	**TF-IDF**	**LDA**
K-means	0.718	0.746	0.448
A-Cluster	0.489	1	0.443
EM	0.383	0.952	0.374

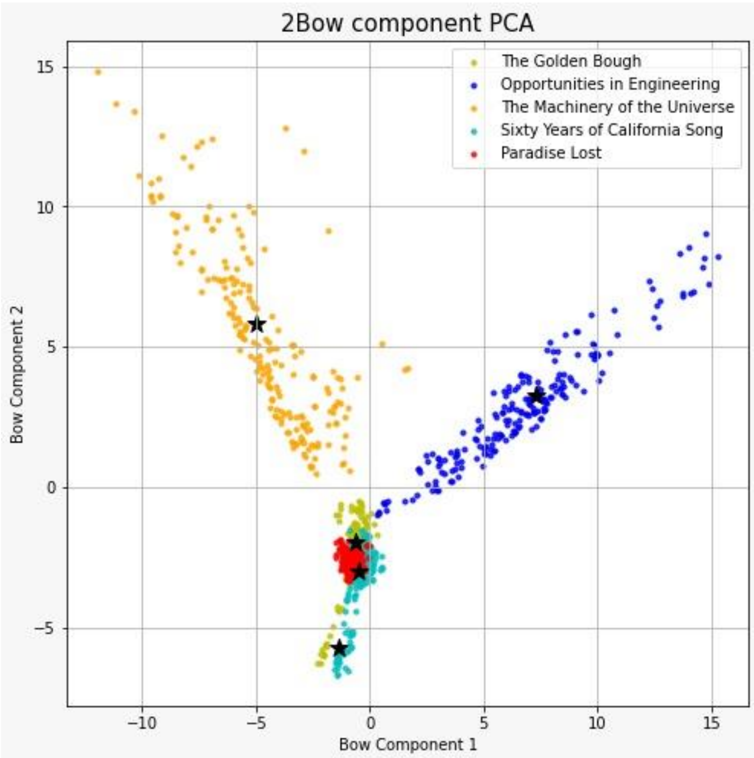


Figure 2: Cluster plot for BOW using K-means

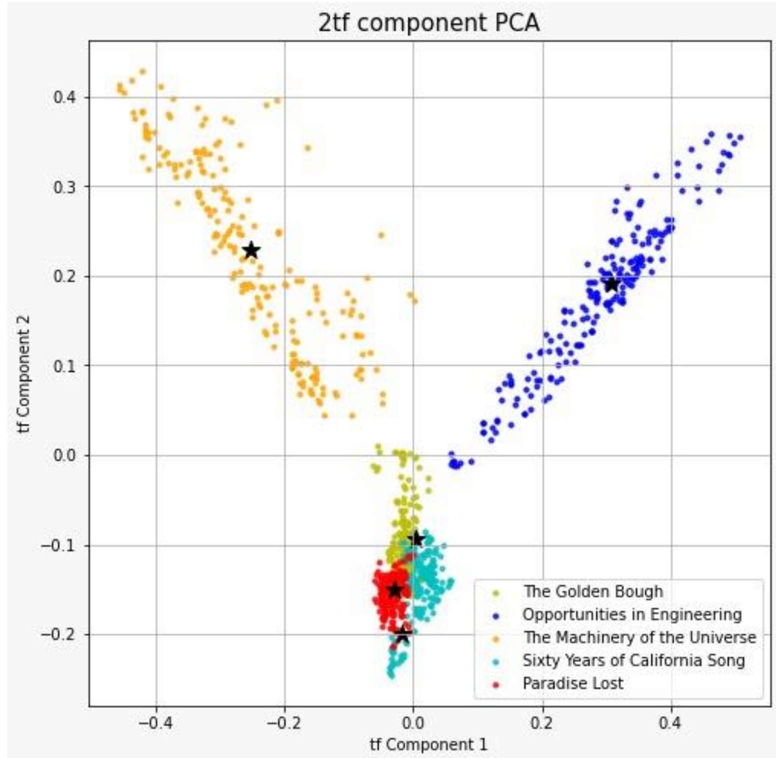
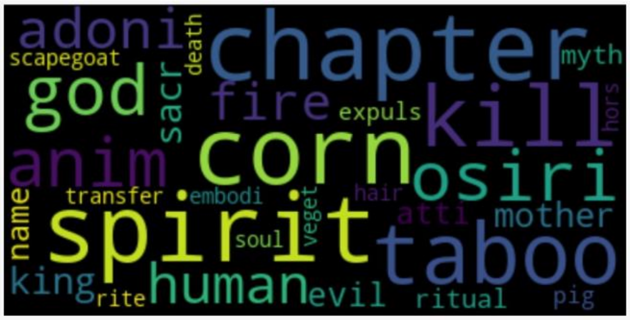
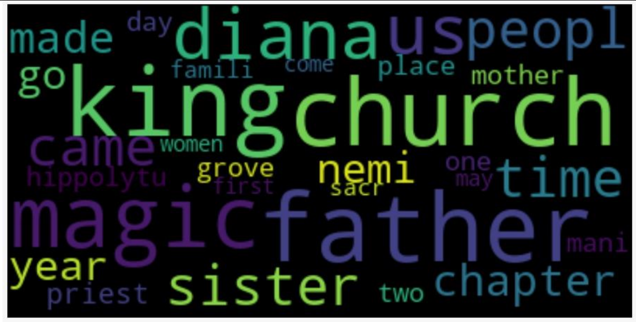
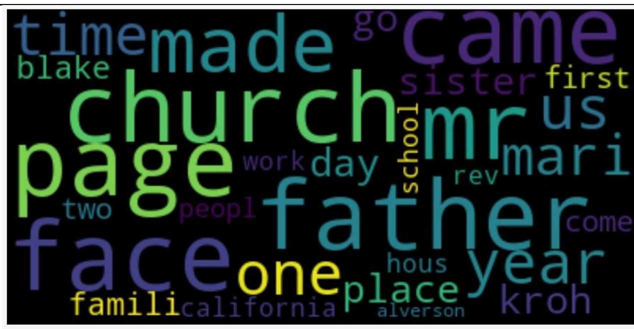
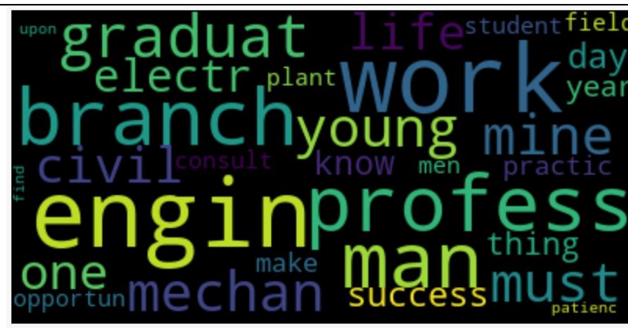
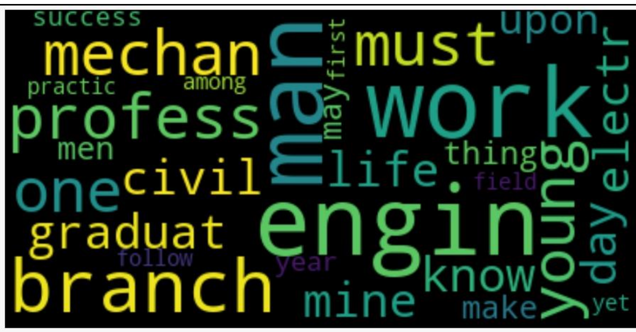
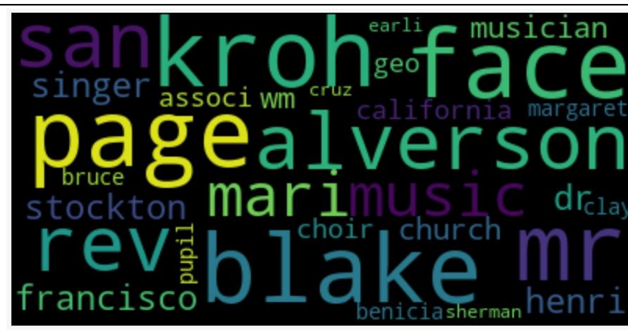
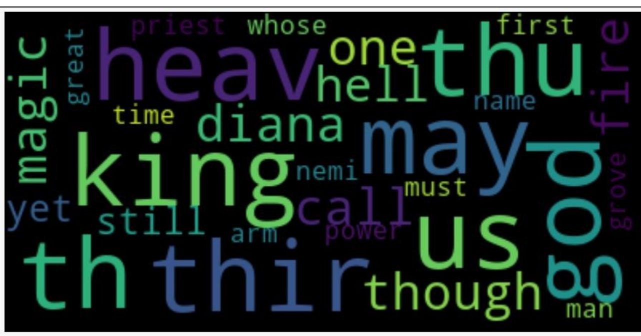
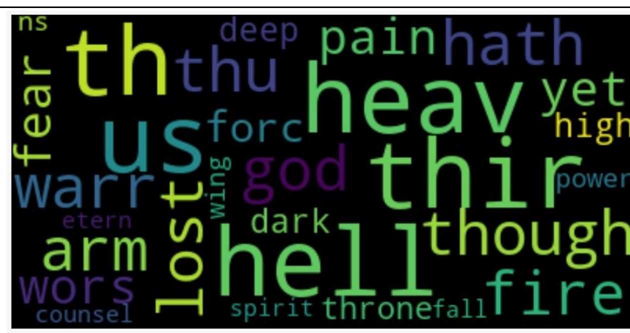
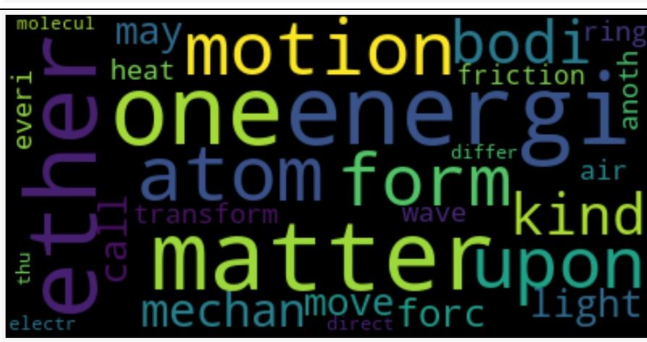
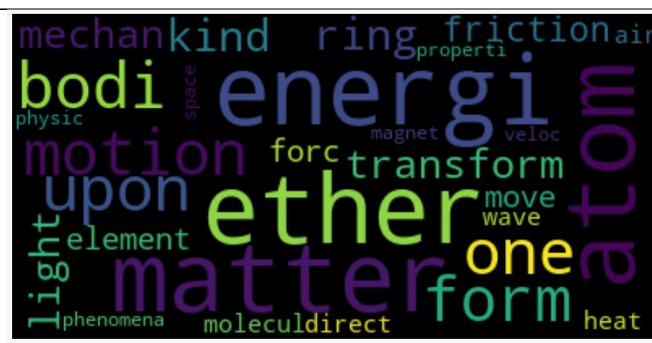


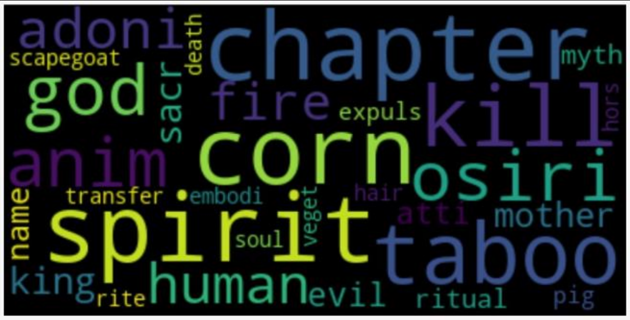
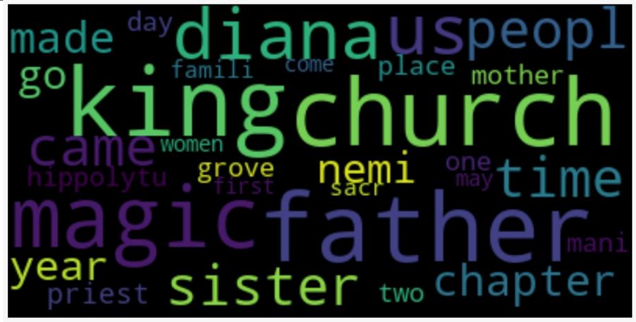
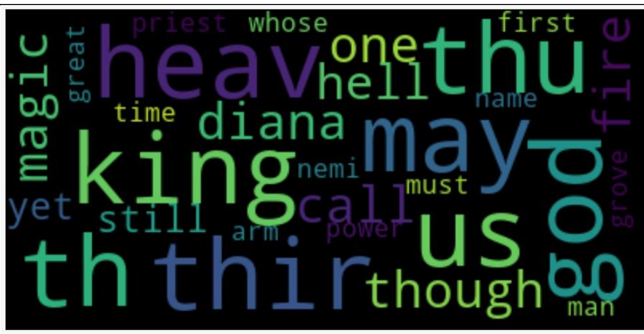
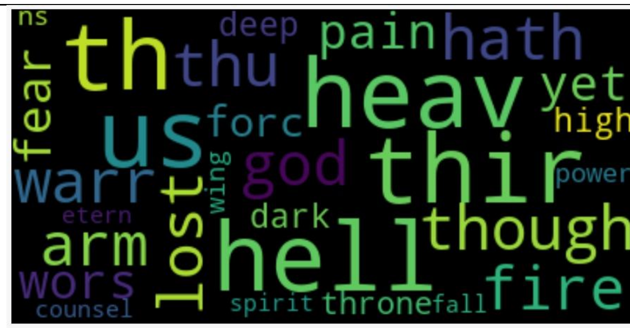
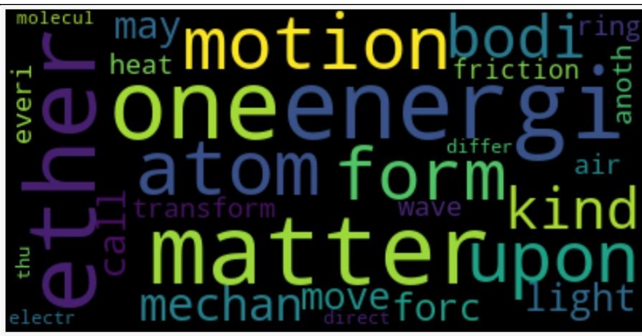
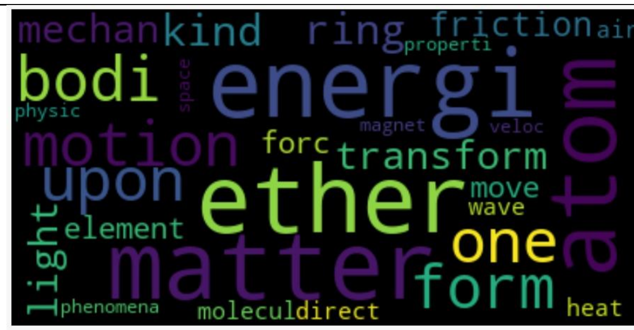
Figure 3: Cluster plot for TF-IDF using K-means

Figure 2 shows the clusters obtained with BOW and K-means. It can be observed that the green, light-blue and red clusters are overlapping, this is an indication that they have similar words between them. The same is observed in Figure 3 which represents the clusters plot for K-means with TF-IDF.

3.4. 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, 3 and 4. From the BOW word cloud table below, we can indeed see similar occurring words between cloud 0,3 and 4. This is one of the reasons why the K-means BOW model is very inaccurate.

BOW word cloud by cluster from 0 to 4	TF-IDF word cloud by cluster from 0 to 4
	
	
	
	
	

<table>
<tr><th>BOW word cloud by cluster 0,3, and 4</th><th>TF-IDF word cloud by cluster 0,3, and 4</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

4. Conclusion

Based on the three clustering models, and the three metrics used. It was found that the model with best Kappa is A-cluster using BOW, the model with best Silhouette is K-means using LDA, and the model with best Rand score is A-cluster using TF-DF. Based on these metrics scores the champion model among those three is the A-cluster with Kappa (0.25), Silhouette (0.06), and Rand Score (1) as we believe that the Rand index is the more effective and more important as it compare the clustering model prediction with the real labels.