NLP assignment:**Clustering**

Ashraqat Torky, Mahmoud Helmy, Sarah Elmasry, Yomna Abdelsattar

June 2021

1 Semantic Analysis: When You Really Want to Understand Meaning in Text

Since the beginning of mankind,Text is at the heart of how we communicate.For humans,the way we understand what's being said is almost an unconscious process. To understand what a text is talking about, we rely on what we already know about language itself and about the concepts present in a text. Machines can't rely on these same techniques.
Semantic analysis describes the process of understanding natural language–the way that humans communicate–based on meaning and context.

2 Introduction about the assignment

2.1 Data Choice

Countries all over the world have different cultures. One of the things that portray these differences is children's books, precisely fairy tale stories. For this assignment, we thought it'd be interesting to analyze different children's fairy tales across different countries. These books are semantically similar because they're fairy tales directed to children. However, each book has its own cultural impact and different genres that it represents. You can imagine princess Mulan in China vs princess Snow White in Europe. Books used:

1. **Japanese Fairy Tales** by Yei Theodora Ozaki (Japan)
2. **Indian Fairy Tales** by Joseph Jacobs and John Dickson Batten (India)
3. **Russian Fairy Tales**: A Choice Collection of Muscovite Folk-lore by Ralston (Russia)
4. **Dutch Fairy Tales** for Young Folks by William Elliot Griffis (Netherlands)
5. **American Fairy Tales** by L. Frank Baum

2.2 Data preparation

Each book is partitioned into 200 random samples, each contains 100 words then Stop words, punctuation, and special characters are eliminated from the data. After looking at the initial results, we removed some extra words that had no useful meaning in the analysis. Such as "yet", "said" and so on.

2.3 Word frequencies with Word Cloud

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.

Conclusion:for the first insight we can see that the words are similar between the 5 different books , so we can say that the kappa and the silhouette are not going to be that highest score .

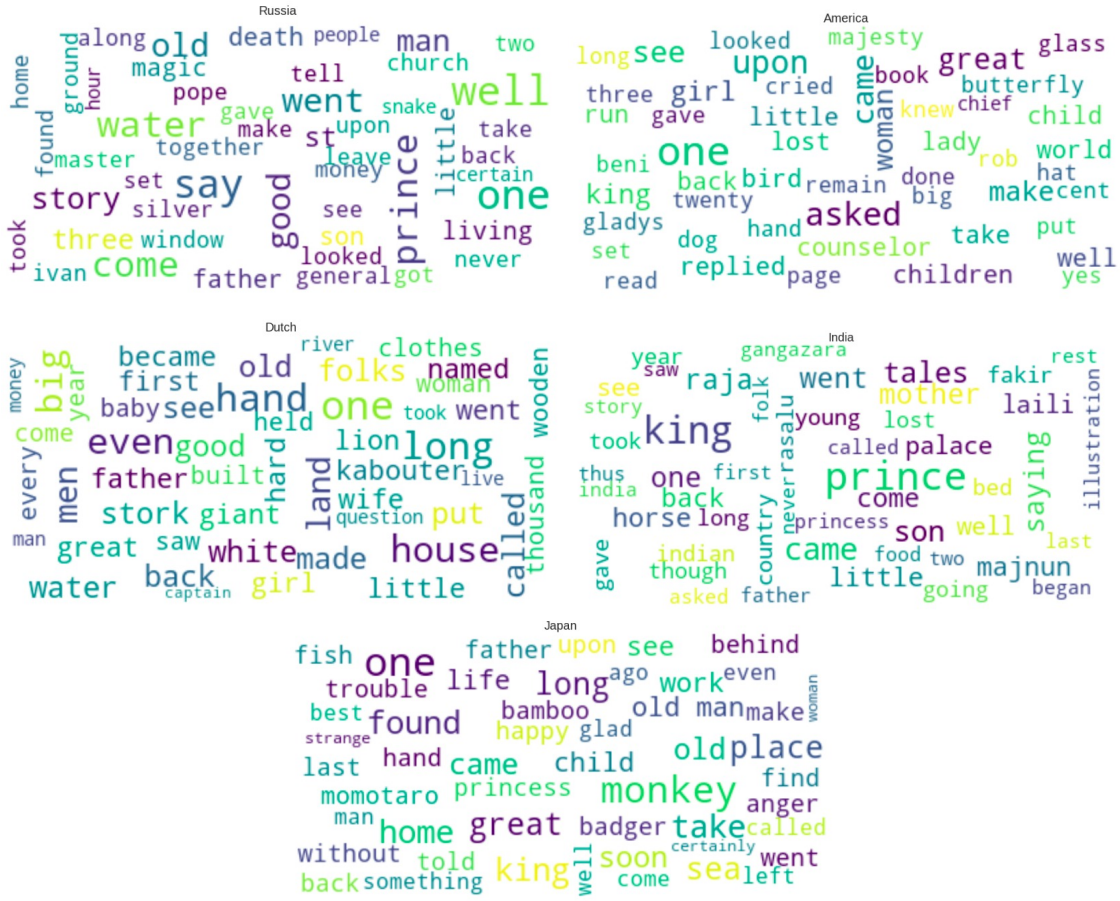

Figure 1: *Word cloud for each document ,**Top left**:The Russian's book version ,**Top right**:The American's book version ,**Bottom left**:The Dutch's book version ,**Bottom right**:The Indian's book version ,**middle Bottom**:The Japanese's book version*

2.3.1 Transformation and Feature Engineering

For feature engineering, different techniques are used: BOW, TF-IDF, n-gram and doc2word

1. **BOW:**
   creates a set of vectors containing the count of word occurrences in the document

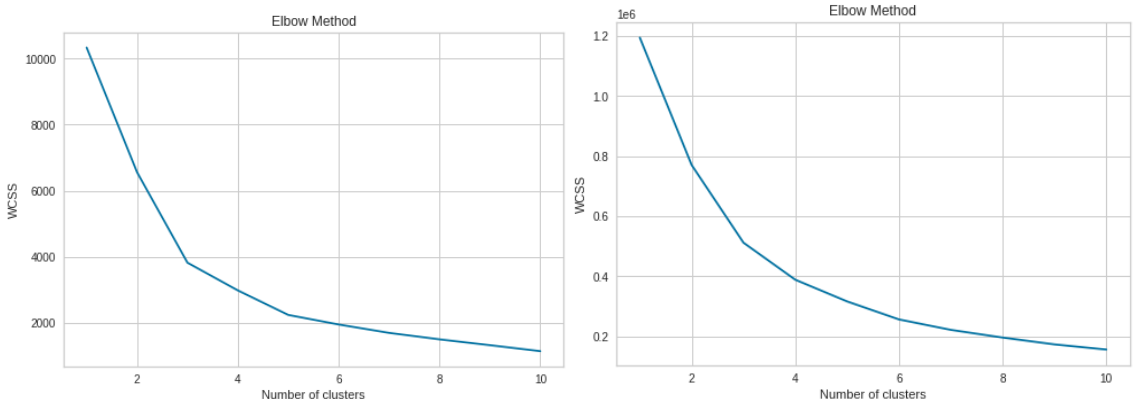

Figure 2: *BOW 's Elbow graphs*

2. **TF-IDF:**
   Contains information on the more important words and the less important ones as well.

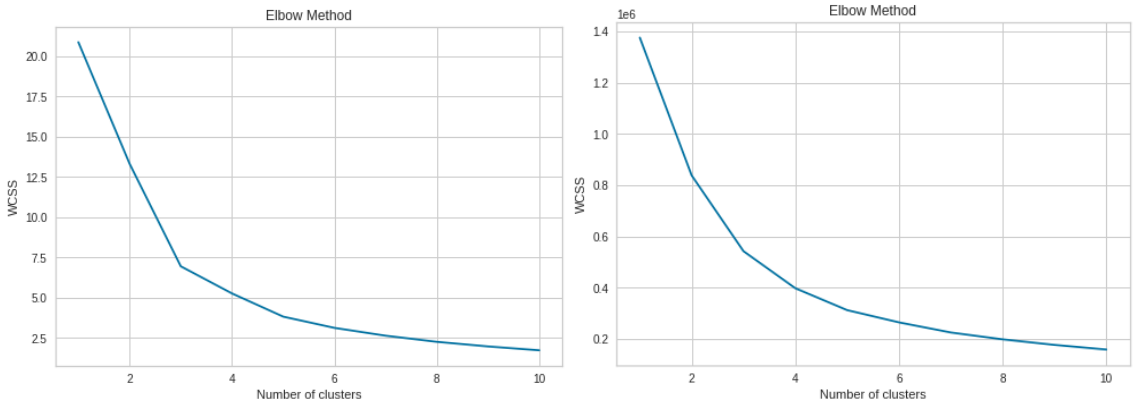

Figure 3: *TF-IDF's Elbow graphs*

3. **n-gram** takes a text variable as input and produces strings corresponding to sliding a window of (user-configurable) n words

   We will not use this feature engineering method because it didn't improve much the results, so we decided to exclude it from the training process

4. **doc2word:**
   represents documents as a vector and is a generalizing of the word2vec method

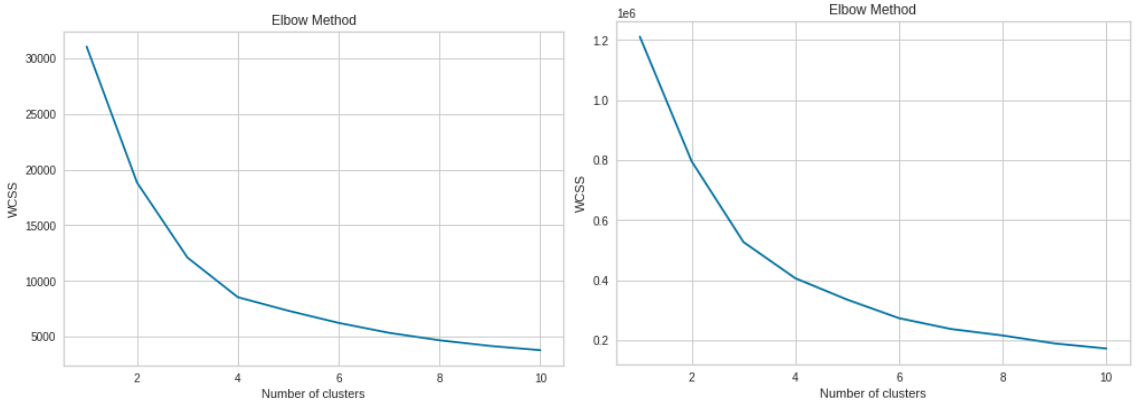

Figure 4: *Doc2word 's Elbow graphs*

Conclusion : From the above graphs, we can conclude that there are no Sharpe elbow in any of the graphs , maybe we can do more feature engineering or excluding more data .
in some graphs,it is clearly appearing the begging of an elbow in k=3 for example . **This is an insight that the model didn't achieve good clustering technique, we are going to investigate to understand more why the missclassification happen**

3 Topic Modeling

Refers to the task of identifying topics that best describes a set of documents. These topics will only emerge during the topic modelling process (therefore called latent).

3.1 Latent Dirichlet Allocation LDA

3.1.1 introduction:

- LDA is the most popular type of topic modeling.
- LDA imagines a fixed set of topics,here we have 5 books so we assume that we have 5 topics. Each topic represents a set of words. And the goal of LDA is to map all the documents to the topics in a way, such that the words in each document are mostly captured by those imaginary topics. We will systematically go through this method by the end which you will be comfortable enough to use this method on your own.

3.1.2 Plots:

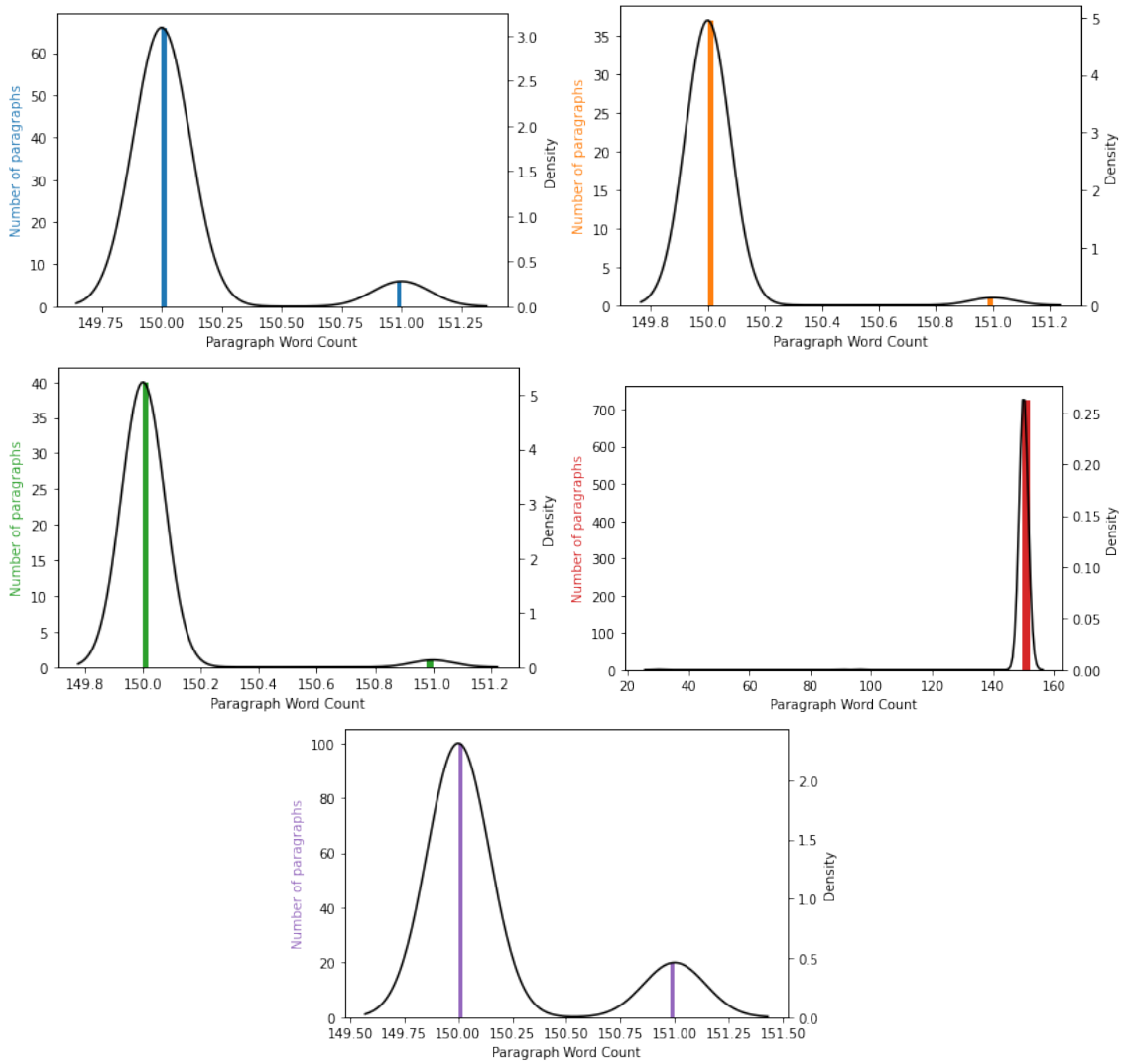

Figure 5: *LDA topic modeling graphs*

Conclusion: From the graphs, we started to conclude that topic 5 and topic 4 are the most frequent among the other topics, because topic 4 comes first because it appears in more than 700 paragraphs and regarding the topic 5 appears more than 100 paragraphs.
This insists on our insights why it is not going to be very hard clustering task for the machine.

4 Clustering algorithms

4.1 Kmeans

K-Means tries to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

4.2 Expectation Maximization (EM)

Expectation Maximization (EM) is a bit more complicated, clustering algorithm that relies on maximizing the likelihood to find the statistical parameters of the underlying sub-populations in the dataset.The EM has high time complexity.

4.3 Agglomerative Hierarchical Clustering

The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity.
Agglomerative clustering works in a "bottom-up" manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes).
For distance Function, we can use Manhattan or Euclidean

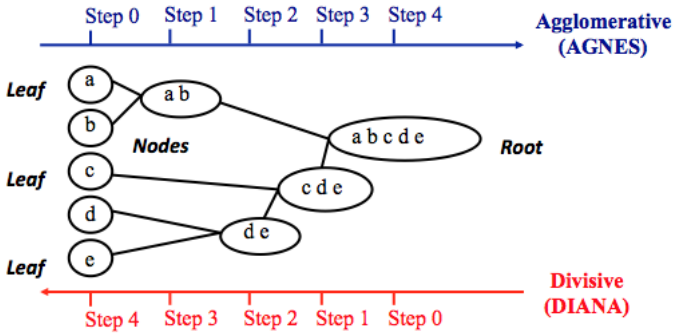

Figure 6: *Agglomerative Clustering process*

5 Evaluation Metrics

5.1 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 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. Negative values indicate that those samples might have been assigned to the wrong cluster.
The silhouette analysis is used to choose an optimal value for number of clusters. The thickness of the silhouette plot the cluster size can be visualized.

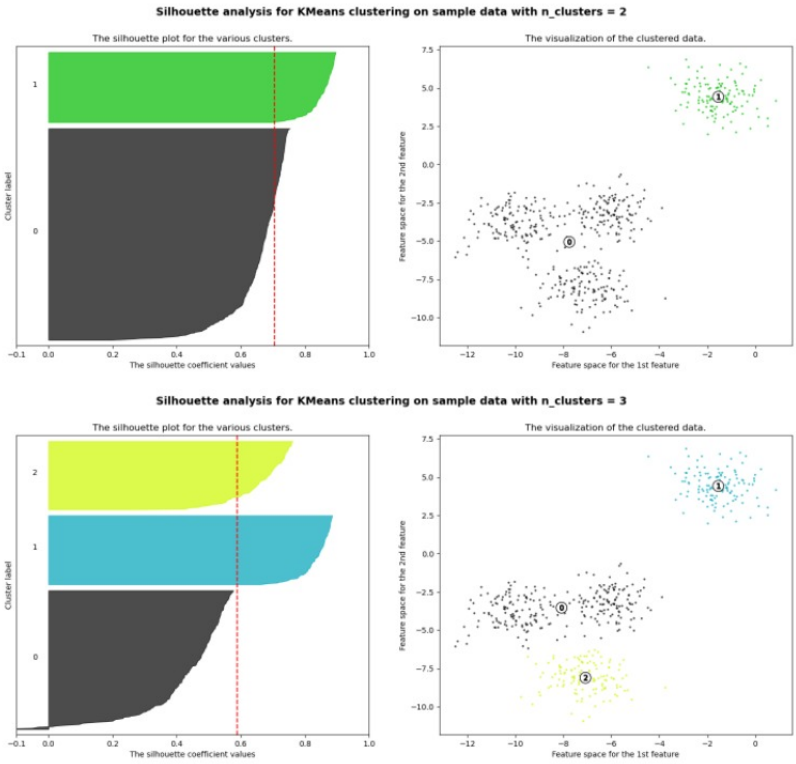

Figure 7: *Silhouette*

5.2 Kappa

Kappa measures the degree of agreement among raters. It gives a score of how much homogeneity there is in the ratings given by annotators.

Cohen's kappa: is always less than or equal to 1. Values of 0 or less, indicate that the classifier is useless. As the value gets closer to 1, it means agreement increases. But there is no standardized way to interpret its values.

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

Figure 8: *Kappa's definition: where po is the observed agreement, and pe is the expected agreement. It basically tells how much better your classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class.*

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}}$$

Figure 9: *Weighted Kappa.*

Fleiss' kappa: a statistical measure of inter-rater reliability for any number of classifiers (multi-voters).

5.3 Coherence

Coherence is an index that basically measures the density, it measures how dense the clustering results are and compare the different cluster results. Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference.
A set of statements or facts is said to be coherent, if they support each other. Thus, a coherent fact set can be interpreted in a context that covers all or most of the facts.

1. **U mass** is based on document co-occurrence counts, a one-preceding segmentation and a logarithmic conditional probability as confirmation measure

2. **C v** measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity

6 Training Process

6.1 BOW

	Error	5 Clusters	3 Clusters
K-means	Kappa	0.1	0.3
	Silhouette	0.001	0.03
Hierarchical Clustering (Euclidean)	Kappa	0	0
	Silhouette	0.004	0.07
Expectation Maximization	Kappa	0.3	0.3
	Silhouette	0.02	0.02

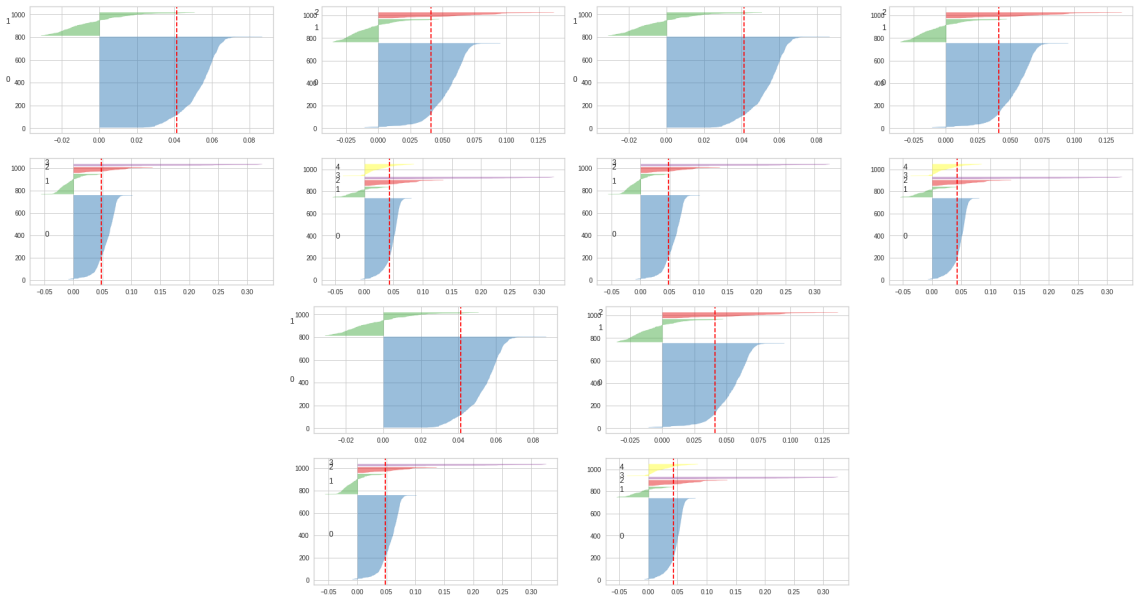

Figure 10: *Silhouette graphs for different algorithms and changing the number of clusters 2,3,4,5* ***Top left:****k-means,* ***Top Right:****Hierarchical clustering,* ***Bottom:****EM*

6.2 TF-IDF

	Error	5 Clusters	3 Clusters
K-means	Kappa	0.4	0.4
	Silhouette	0.01	0.027
Hierarchical Clustering(Manhattan)	Kappa	0.04	0.4
	Silhouette	0.006	0.03
Expectation Maximization	Kappa	0	0.1
	Silhouette	0.001	0.013

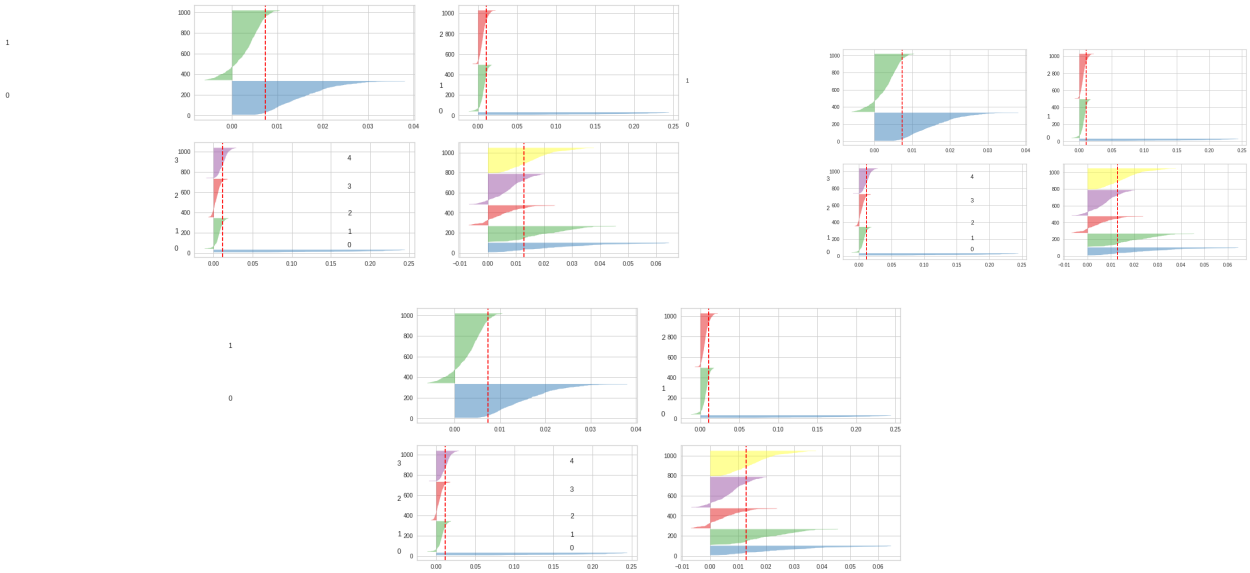

Figure 11: *Silhouette graphs for different algorithms and changing the number of clusters 2,3,4,5* ***Top left:****k-means,* ***Top Right:****Hierarchical clustering,* ***Bottom:****EM*

6.3 Doc2word

	Error	5 Clusters	3 Clusters
K-means	Kappa	0.5	0.6
	Silhouette	-0.001	0.001
Hierarchical Clustering (Euclidean)	Kappa	0.5	0.6
	Silhouette	-0.008	0.02
Expectation Maximization	Kappa	0.5	0
	Silhouette	-0.01	0.06

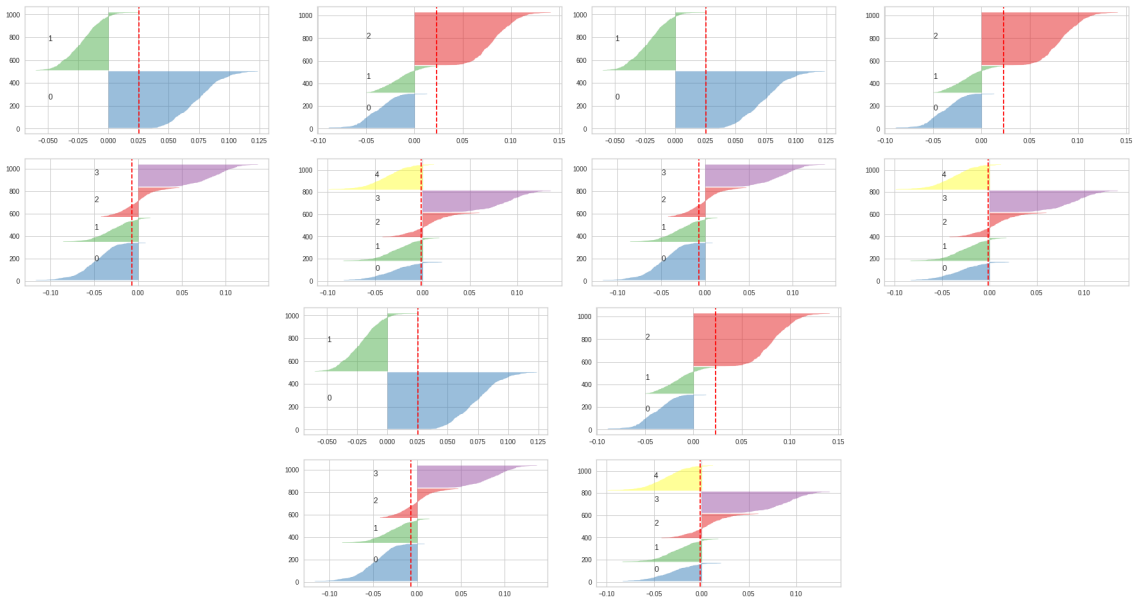

Figure 12: *Silhouette graphs for different algorithms and changing the number of clusters 2,3,4,5* ***Top left:****k-means,****Top Right:****Hierarchical clustering,****Bottom:****EM*

General Conclusion Silhouette values are generally low because our sample books are semantically similar so there are samples that are near the boundary between two neighboring clusters. But hierarchical clustering has the best silhouette value here.

6.4 Champion Model

We notice that best kappa values are in doc2Word k-means and doc2Word hierarchical, but because the silhouette value of hierarchical is higher than k-means, we choose the doc2Word hierarchical clustering as our champion model. If our choice criteria is silhouette value, then our champion model would be Kmeans with feature engineering Doc2Word

7 Error Analysis:

7.1 K-means

During the training in K-means it was concluded on the same dataset with k=3 and k=5 ,it is clearly that the model is conflicting between 4 books, 2 books in each cluster. There are a huge overlap between the books due to the similarity with each other

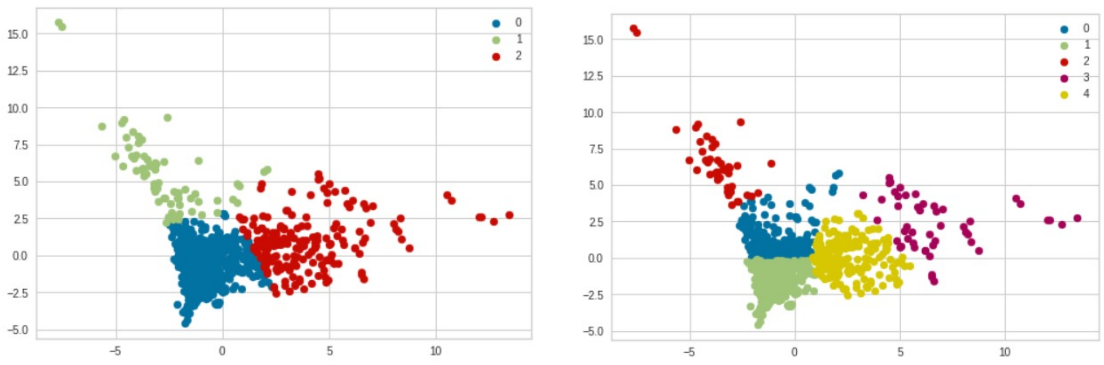

Figure 13: *plotting k-means graphs using 3 clusters and 5 clusters*

7.2 Frequent's Words Topics

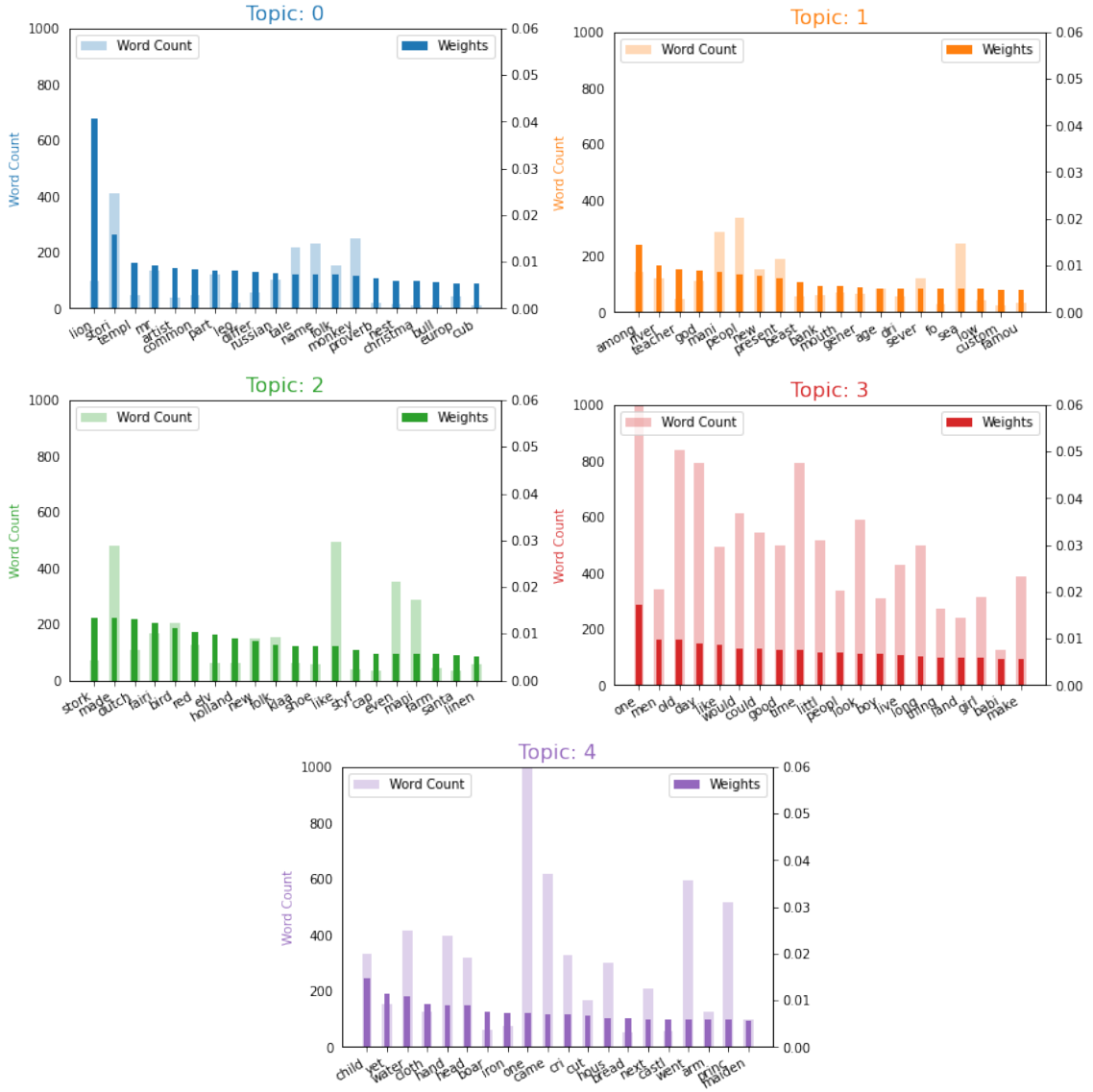

Figure 14: *Frequents words in each topic*

Conclusion: from this frequencies we can visualize why the model miss-classify many topics.

7.3 Word Cloud Topics

Figure 15: *Word Cloud in each topic with LDA*

Conclusion: Comparing with the normal word cloud, we can see that the model took frequent word in one topic like monkey were classified in the first with japan but after Lda it was wrongly assigned to a cluster with Russian and lion , from these plot it is obvious that the model didn't cluster the topic well. The model missclassified some segments from the books due to similar frequent words as:

- "old", in Russian and Dutch
- "king", in American and Indian
- "father" in Russian and Japanese On the other hand there were unique words that distinct segments as:
- "raja majun" in Indian
- "momotaro" in Japanese

8 references

- K-means and Expectation Maximization (EM)
  https://radiant-brushlands-42789.herokuapp.com/towardsdatascience.com/a-comparison-between-k-
- Agglomerative Hierarchical Clustring
  https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/
- Plotting
  https://www.machinelearningplus.com/topic-modeling-visualization-how-to-present-results-lda-
  #4.-Build-the-Bigram,-Trigram-Models-and-Lemmatize
  https://matplotlib.org
- Silhouette
  https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_
  analysis.html
- Coherence
  https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-ld
- Kappa
  https://towardsdatascience.com/inter-rater-agreement-kappas-69cd8b91ff75
- LDA
  https://radiant-brushlands-42789.herokuapp.com/towardsdatascience.com/light-on-math-machine-
- clean data frame using regex
  https://github.com/karolzak/support-tickets-classification#22-dataset
- Japanese Fairy Tales by Yei Theodora Ozaki
  https://www.gutenberg.org/ebooks/4018
- Indian Fairy Tales by Joseph Jacobs and John Dickson Batten
  https://www.gutenberg.org/ebooks/7128
- Russian Fairy Tales: A Choice Collection of Muscovite Folk-lore by Ralston
  https://www.gutenberg.org/ebooks/22373
- Dutch Fairy Tales for Young Folks by William Elliot Griffis
  https://www.gutenberg.org/ebooks/7871
- American Fairy Tales by L. Frank Baum
  https://www.gutenberg.org/ebooks/4357