[DTI5125 [EG] Data Science Applications
Group 10, Assignment 2

Text Clustering

Names:

Rim Mahmoud Tony

Nada Mohamed Zakaria

Dina Ibrahim Mohammady

Sema Abdel Nasser Mosaad

Abstract

Text clustering is a method used in machine learning and natural language processing to group together comparable texts or documents based on their content. It is an unsupervised learning method that seeks to categorize meaningfully unstructured text data. Text clustering can be done using a variety of techniques, including as hierarchical clustering, k-means clustering, and density-based clustering. Text clustering is useful in many domains and can help to find insights and patterns in very big and very different datasets.

Business case and Dataset

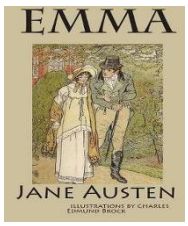
The books we selected, including "Emma" by Jane Austen, "Leaves of Grass" by Walt Whitman, "Paradise Lost" by John Milton, "Moby-Dick; or, The Whale" by Herman Melville, and "The Man Who Was Thursday" by G.K. Chesterton, though that each book belongs to different author and different genre but all of them is a classics of their respective genres and have had a significant influence on literary culture

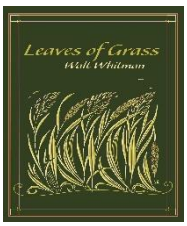
Although the themes and subject matter of these books vary, they all place a strong emphasis on the value of language, the complexity of the human condition, and the endurance of art. However, because these books all roughly date from the same period, it

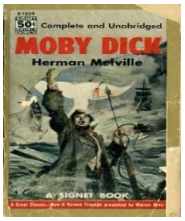
can occasionally be challenging to spot trends and group similar texts together based on their content.by means of customary

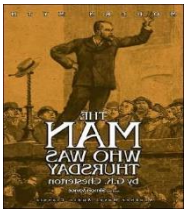
It can be challenging to spot trends and put related texts together based on their content because almost all of these books are from the same era .through means of analysis that is customary Consequently, text clustering can be an effective technique for studying and comprehending enormous quantities of literature.

Books









Preprocessing and Data Cleansing

- Listing all the books in Gutenberg's library
- Choose five different books by five different authors and different genre
- Data preparation:
    I. Removing stop words
    II. Converting all words to the lower case
    III. Tokenize the text
    IV. Lemmatization is the next step that reduces a word to its base form
- Data Partitioning: partition each book into 200 documents, each document is a 150 word record
- Data labeling as follows:
    I. austen-emma→ a
    II. whitman-leaves → b
    III. milton-paradise→ c
    IV. melville-moby_dick→ d
    V. chesterton-thursday→ e

	Partitions	Book Name	Book Label	Book Author
0	forced give attention father sister nothing wo...	Emma.txt	a	Jane Austen
1	christmas day need find excuse elton absenting...	Emma.txt	a	Jane Austen
2	welcomed utmost delight father trembling dange...	Emma.txt	a	Jane Austen
3	far shall call upon miss bates still nearer kn...	Emma.txt	a	Jane Austen
4	emma said something civil excellence miss fair...	Emma.txt	a	Jane Austen
...	...	...	...	...
995	garden empty everyone gone long ago went rathe...	Thursday.txt	e	G. K. Chesterton
996	safely paris might try kidnap lock well known ...	Thursday.txt	e	G. K. Chesterton
997	prof good natured trick read bible part laugh ...	Thursday.txt	e	G. K. Chesterton
998	protected eye still impenetrable professor sig...	Thursday.txt	e	G. K. Chesterton
999	playing fool hour peril make syme think comrad...	Thursday.txt	e	G. K. Chesterton

1000 rows × 4 columns

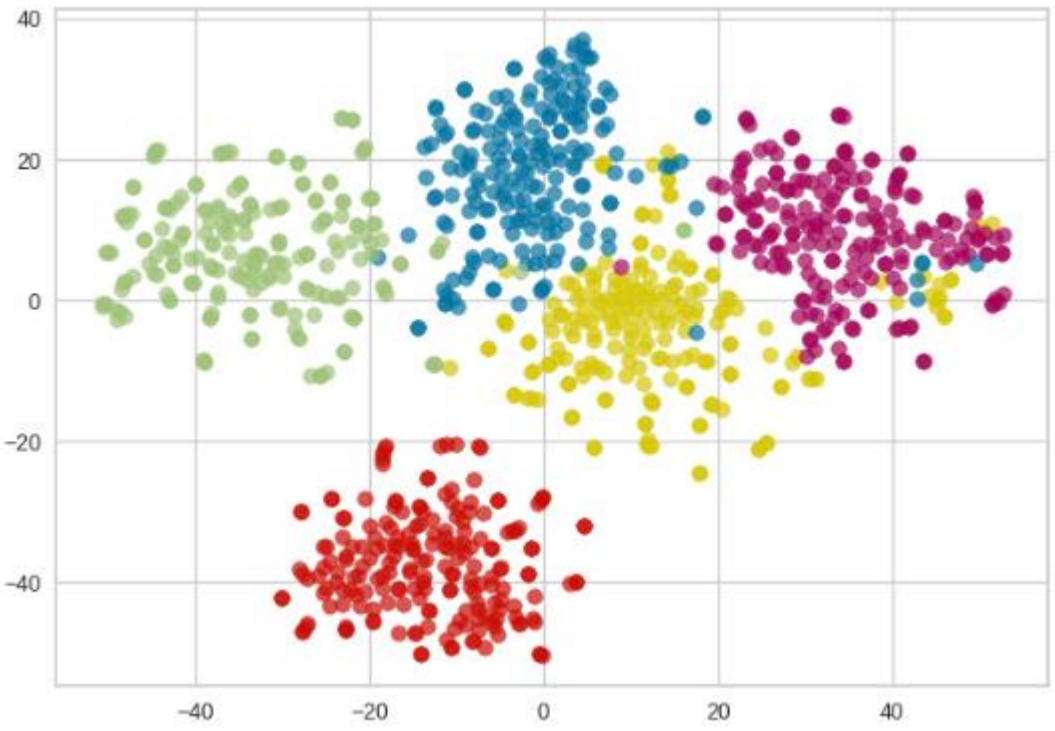
Feature Engineering

- Encoding

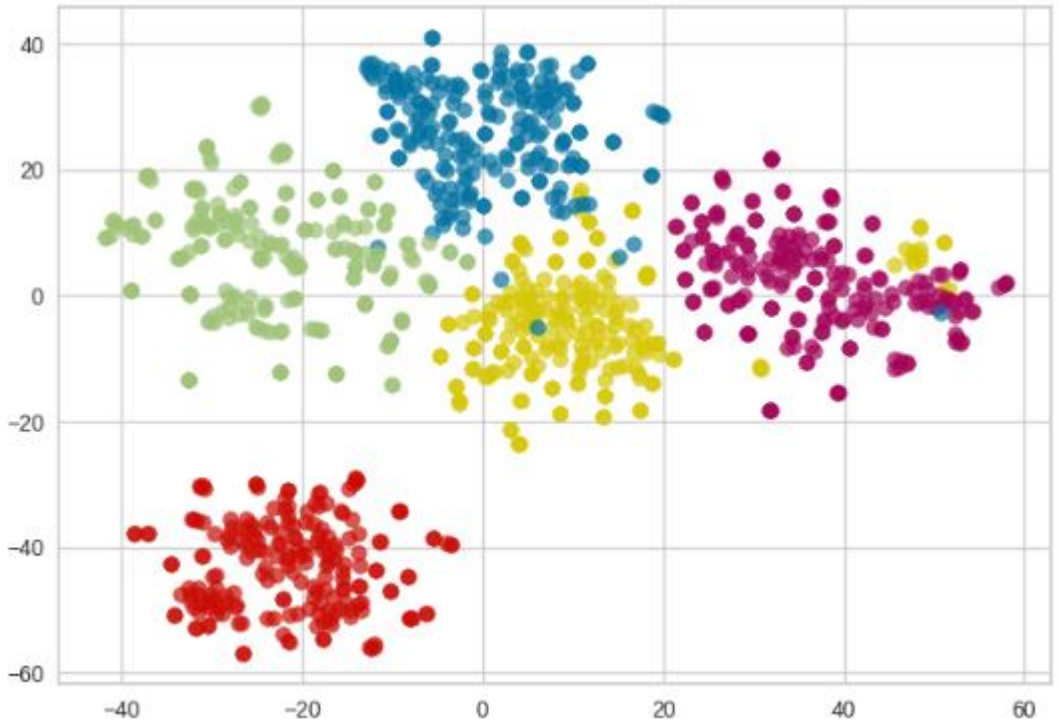
- Text transformation using BOW, TF-IDF , LDA , Word-Embedding

  • **BOW :** It represents the occurrence of words within a document, it involves two things:

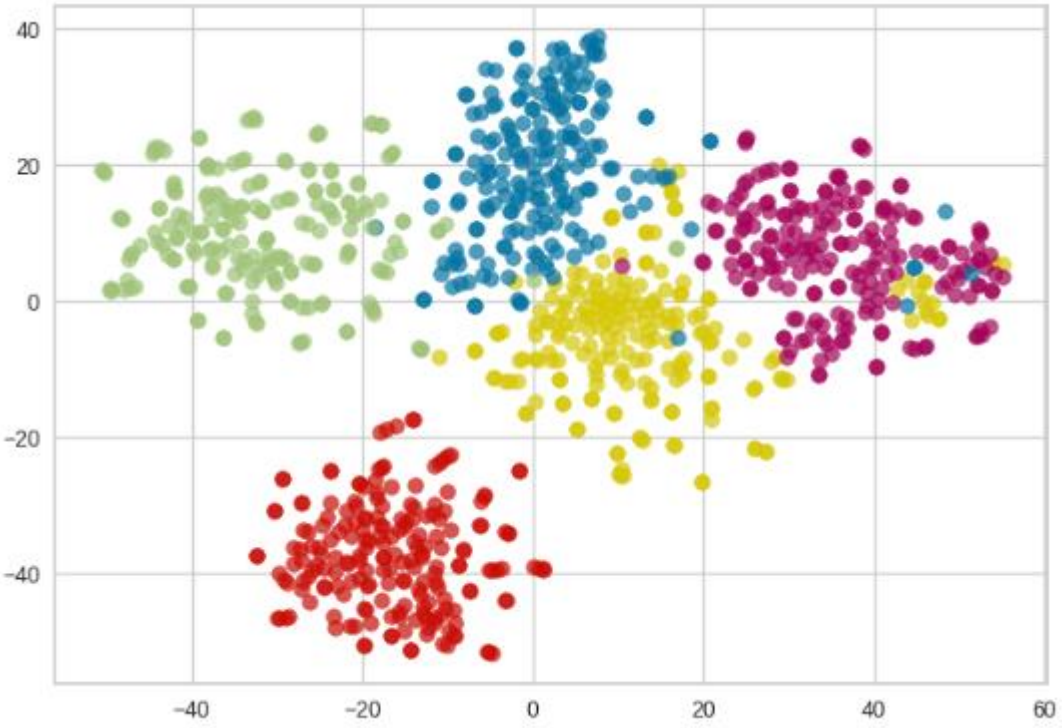
    - A vocabulary of known words
    - A measure of the presence of known words.



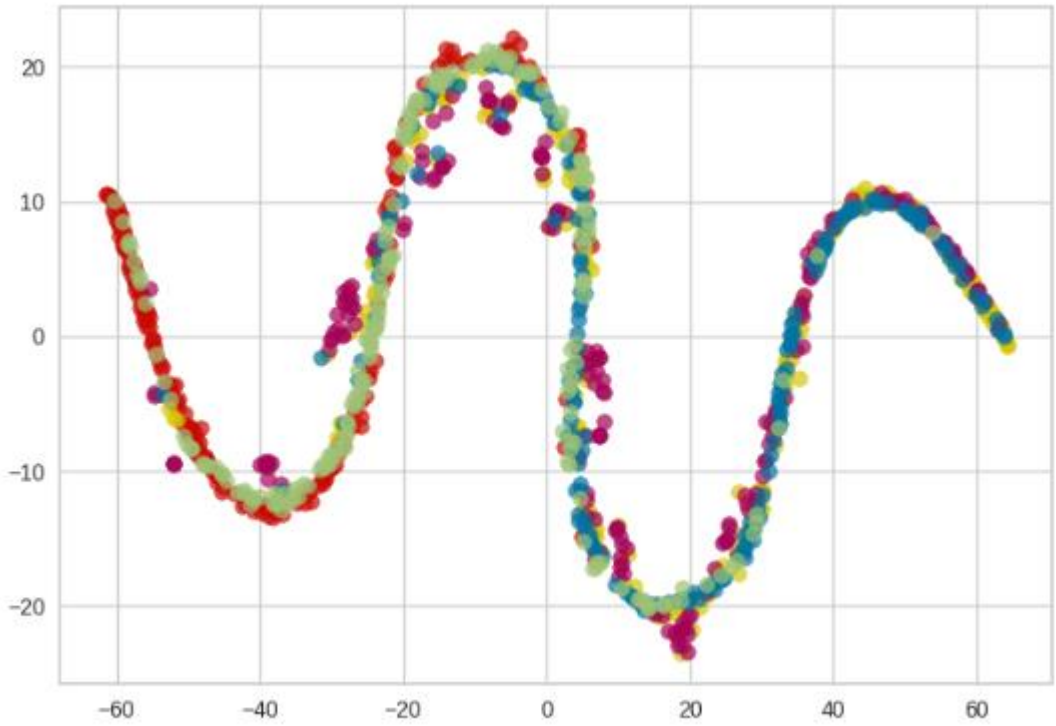
• **TF-IDF:** a technique to quantify words in a set of documents. We compute a score for each word to signify its importance in the document and corpus.



• **LDA:** (Latent Dirichlet Allocation) is a generative statistical model used for topic modeling, a technique that uncovers hidden thematic structures within a collection of documents. LDA assumes that each document is a mixture of topics, and each topic is a distribution of words.



• **Word embedding:** is a technique used in natural language processing to represent words or phrases as dense and continuous vectors in a high-dimensional space. It aims to capture the semantic and contextual relationships between words, enabling machines to understand and process human language more effectively.



Dimensionality Reduction After text transformation, it is quite hard to deal with this huge number of patterns and here dimensionality reduction comes to just provide us with the most important features without any redundancy. PCA and T-SNE are good choices to use in such situations. T-SNE differs from PCA by preserving only small pairwise distances or local similarities whereas PCA is concerned with preserving large pairwise distances to maximize variance, so we decide to go with T-SNE as a pre step to the modelling.

Benefits of using T-SNE:

- Reduce dimensions of the data.
- Reduce computation time.
- Good visualizations.

clustering algorithms

For each technique of the above, these following models are used.

1. K-Means.

2. Expectation Maximization (EM).

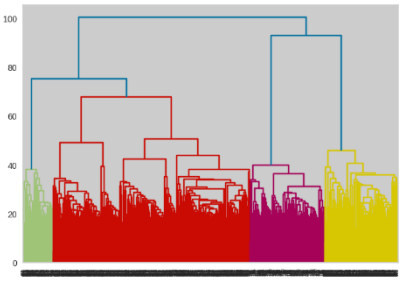
3. Hierarchical clustering (Agglomerative).

1. K-Means is one of the simplest and most popular machine learning algorithms out there. It is an unsupervised algorithm as it doesn't use labelled data, in our case it means that no single text belongs to a class or group.

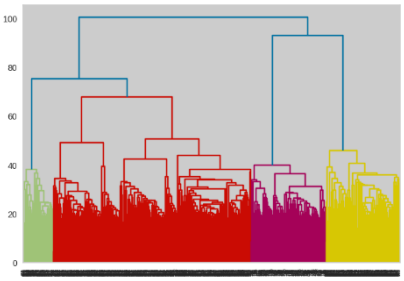
2. EM is a technique for carrying out maximum likelihood estimation when latent variables are present. To achieve this, the model is first optimized , then the latent variable values are estimated, and so on until convergence

3. Hierarchical clustering dictates which distance to employ between sets of observation, and we used ward to minimise the variance of the clusters being merged. Euclidean metrics are used to compute the linkage

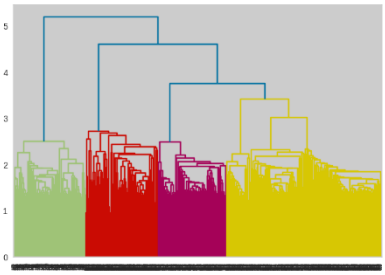
Hierarchical clustering With BOW



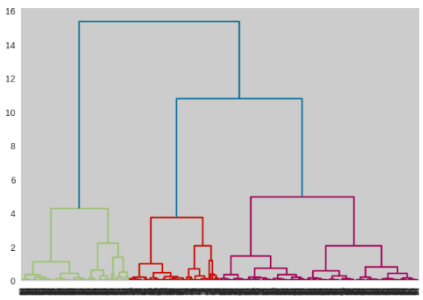
Hierarchical clustering With LDA



Hierarchical clustering With TF



Hierarchical clustering With Word Embedding



Model Evaluation

We Evaluate our models using different metrics as:

1. Silhouette
2. Cohen’s kappa
3. Coherence With LDA

Silhouette

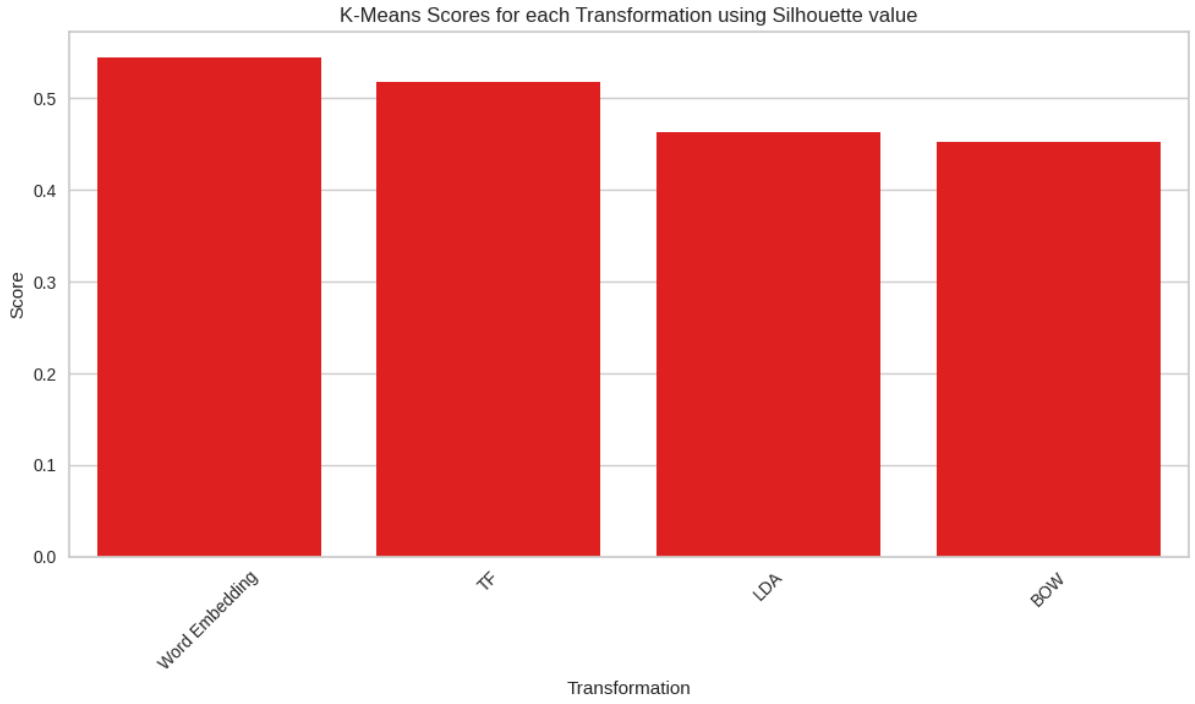
Silhouette DataFrame:

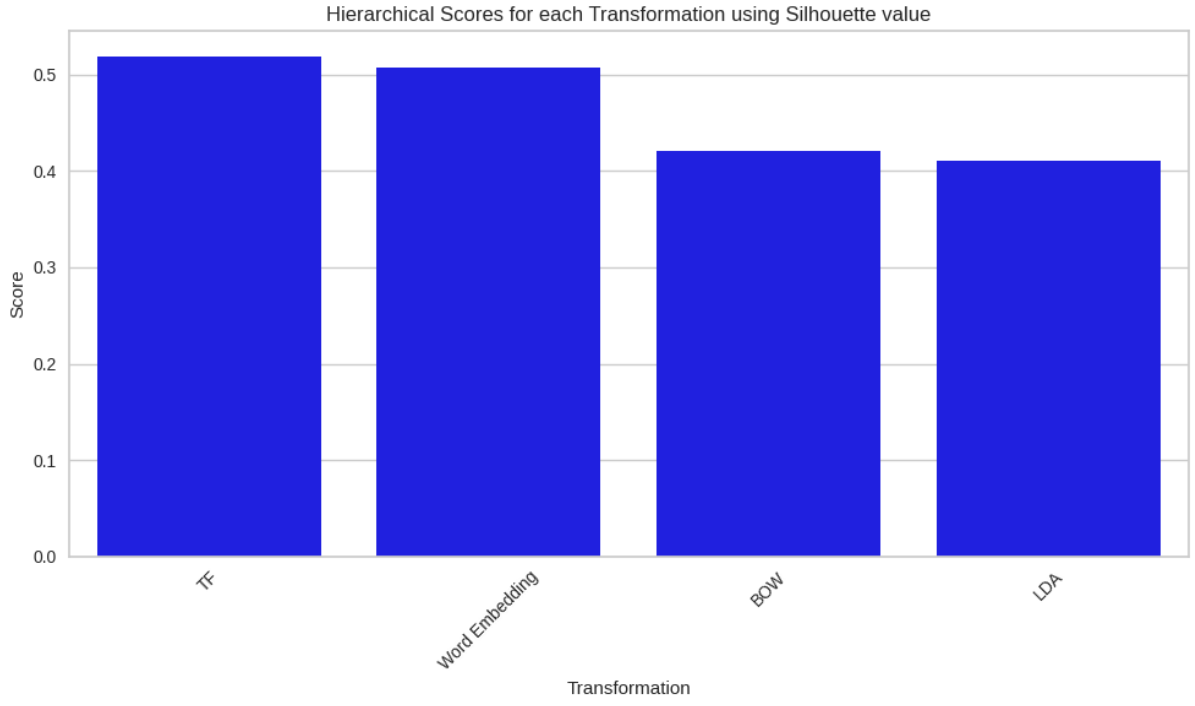
	Model	Transformation	Score
0	K-Means	Word Embedding	0.544873
1	Hierarchical	TF	0.519120
2	K-Means	TF	0.517842
3	EM	TF	0.511162
4	Hierarchical	Word Embedding	0.506720
5	EM	Word Embedding	0.486029
6	K-Means	LDA	0.463269
7	EM	BOW	0.463024
8	EM	LDA	0.457803
9	K-Means	BOW	0.453050
10	Hierarchical	BOW	0.420360
11	Hierarchical	LDA	0.410231

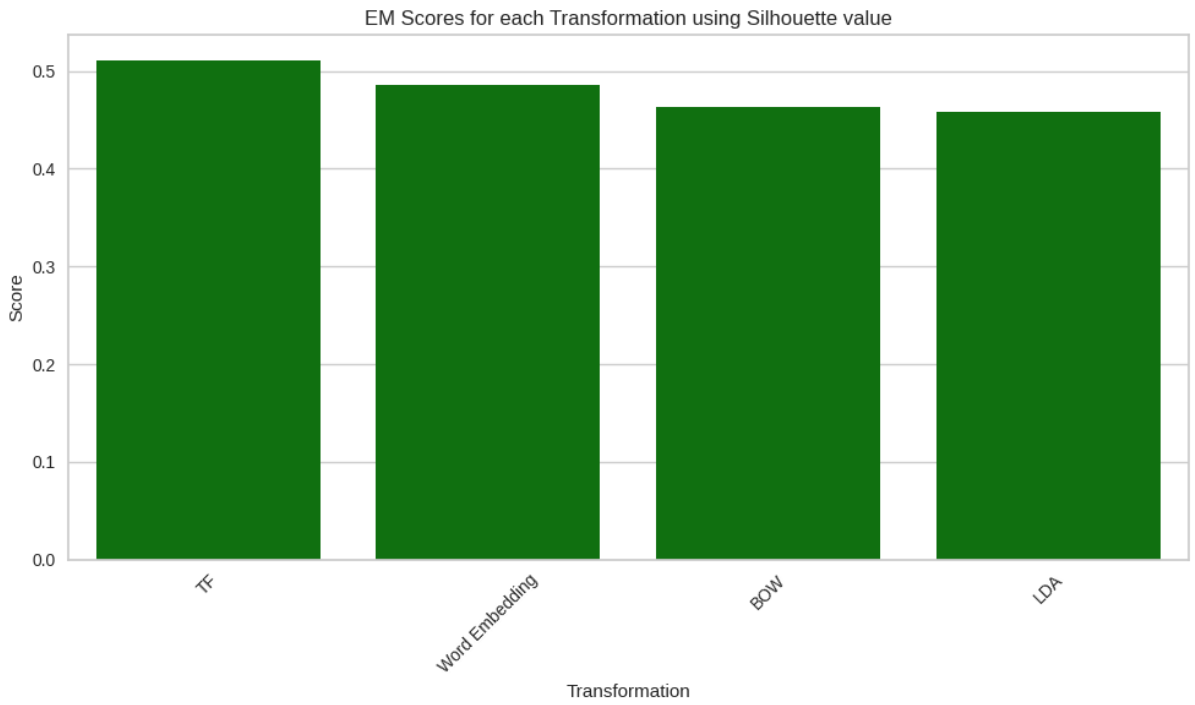
Cohen's Kappa

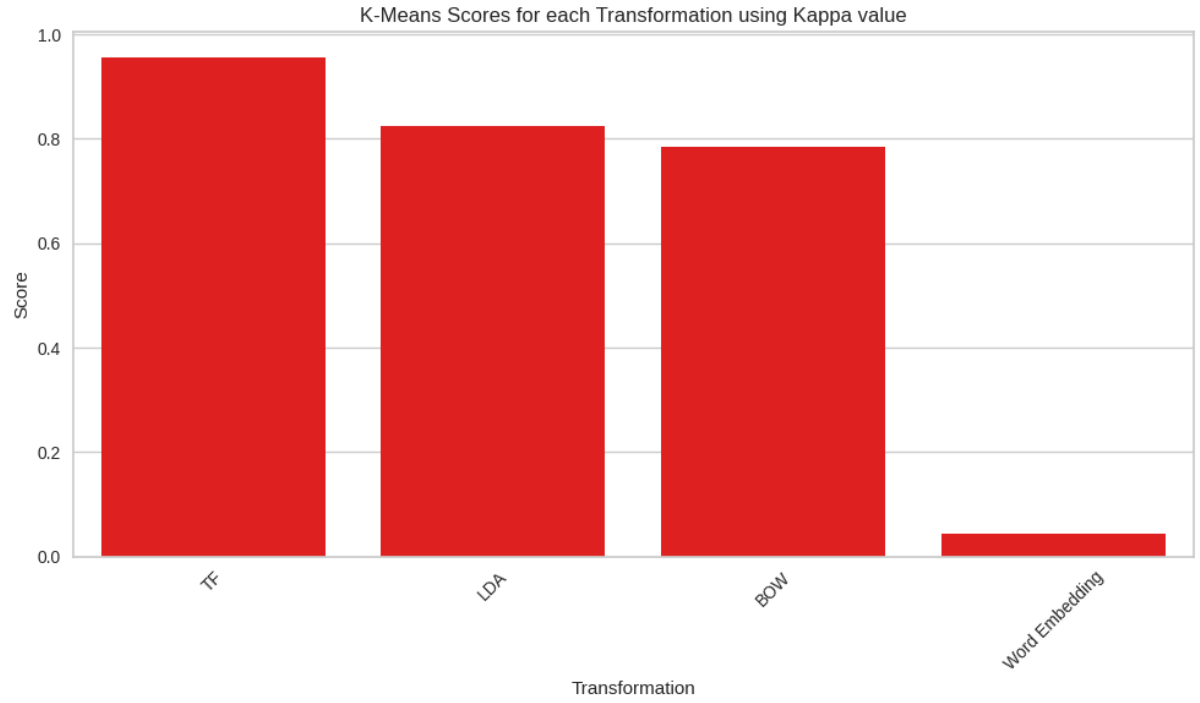
Kappa DataFrame:

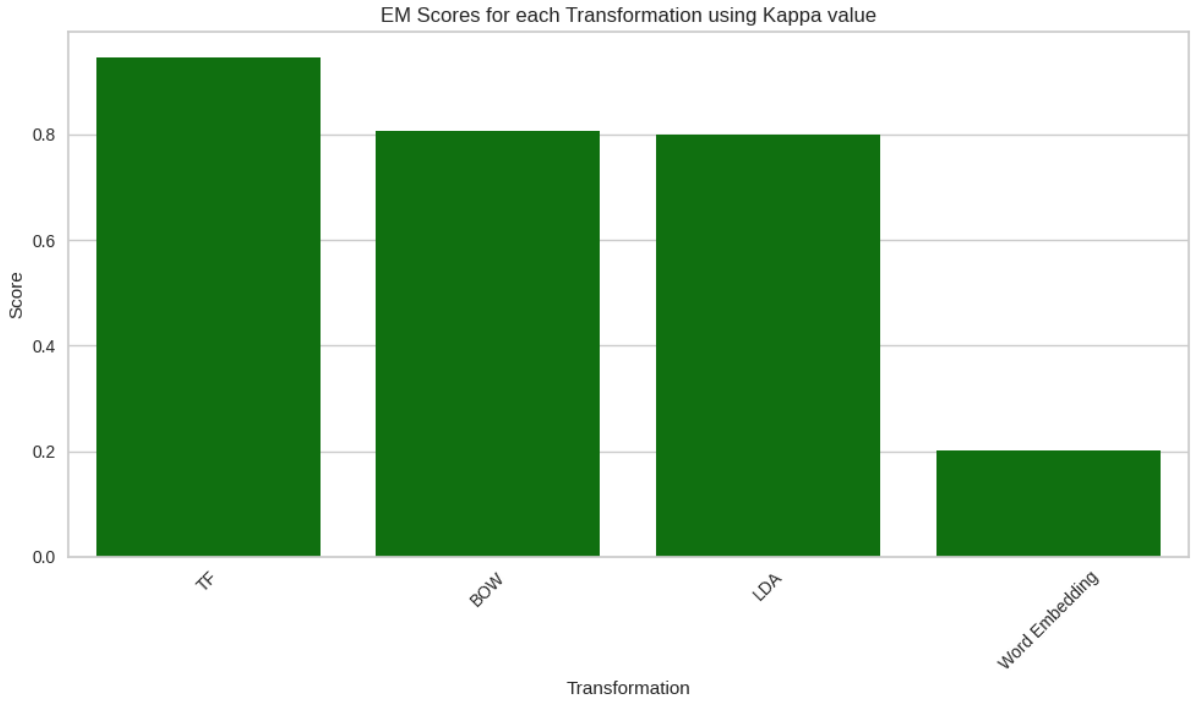
	Model	Transformation	Score
0	K-Means	TF	0.956250
1	EM	TF	0.946250
2	Hierarchical	TF	0.933750
3	K-Means	LDA	0.825000
4	Hierarchical	LDA	0.808750
5	EM	BOW	0.806250
6	EM	LDA	0.798750
7	K-Means	BOW	0.783750
8	Hierarchical	BOW	0.726250
9	EM	Word Embedding	0.201250
10	Hierarchical	Word Embedding	0.197500
11	K-Means	Word Embedding	0.045000

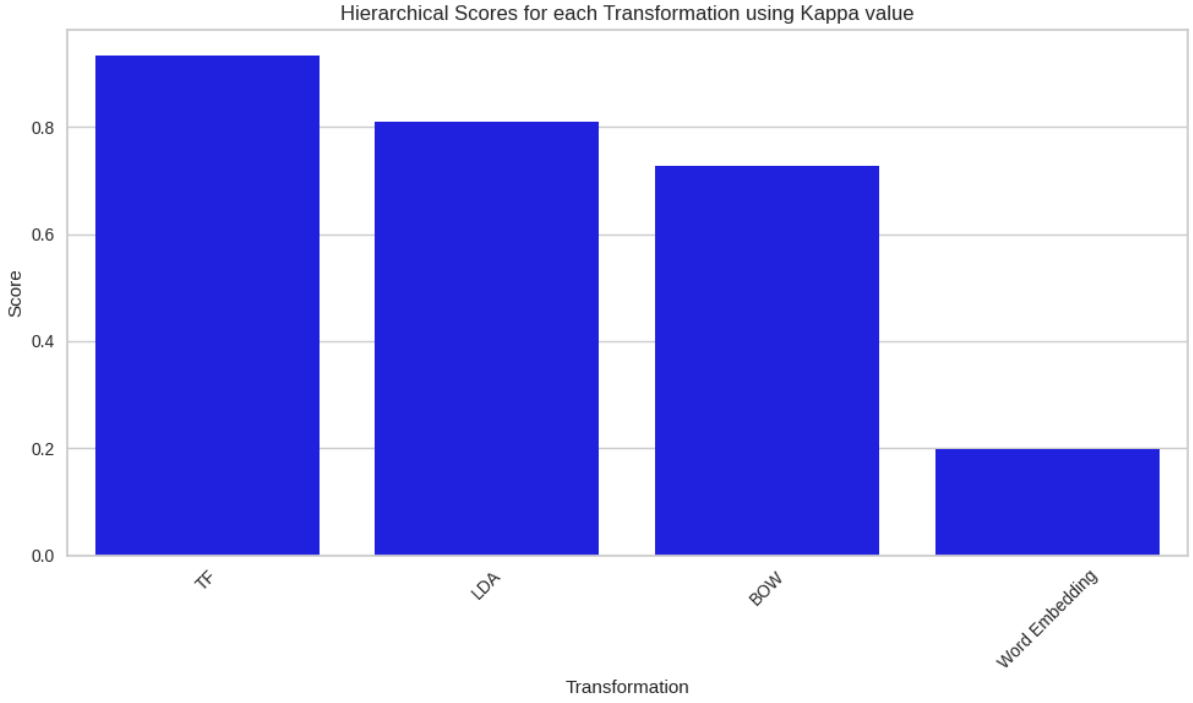

K-Means Scores for each Transformation using Silhouette value
Score
0.0
0.1
0.2
0.3
0.4
0.5
Word Embedding
TF
LDA
BOW
Transformation

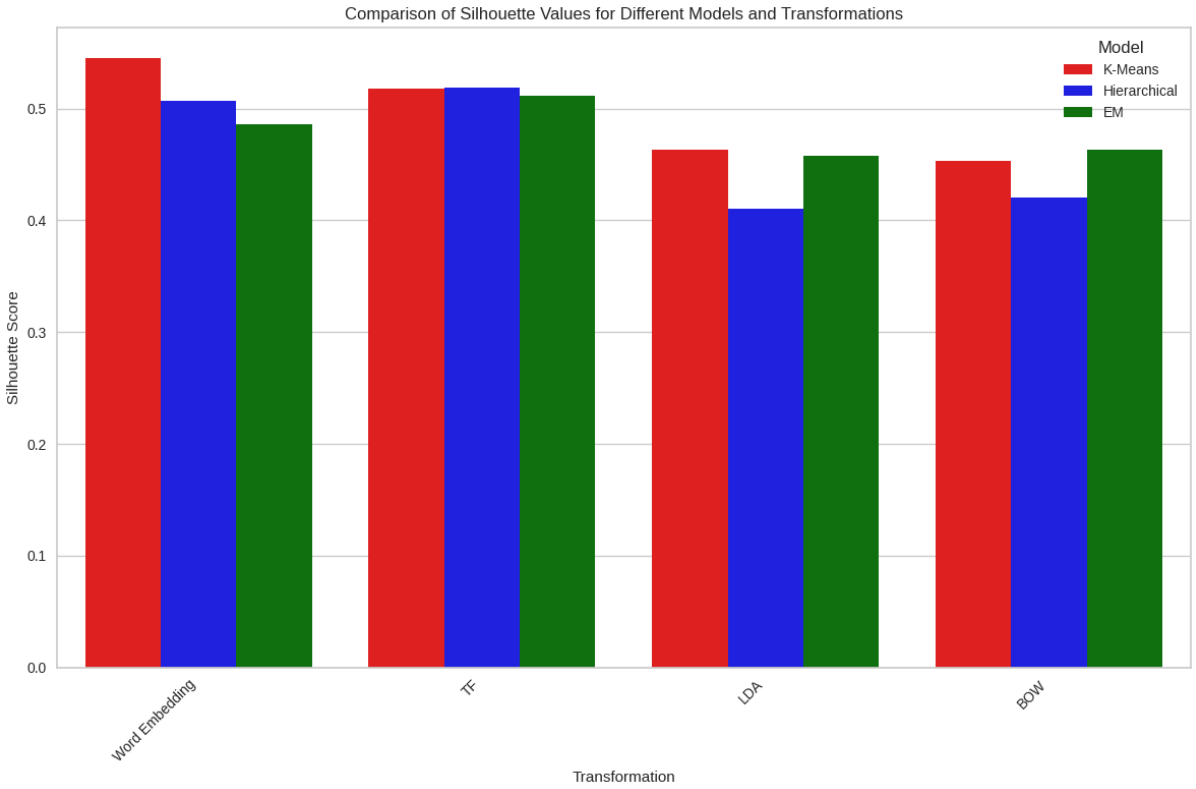

Hierarchical Scores for each Transformation using Silhouette value
Score
0.0
0.1
0.2
0.3
0.4
0.5
TF
Word Embedding
BOW
LDA
Transformation

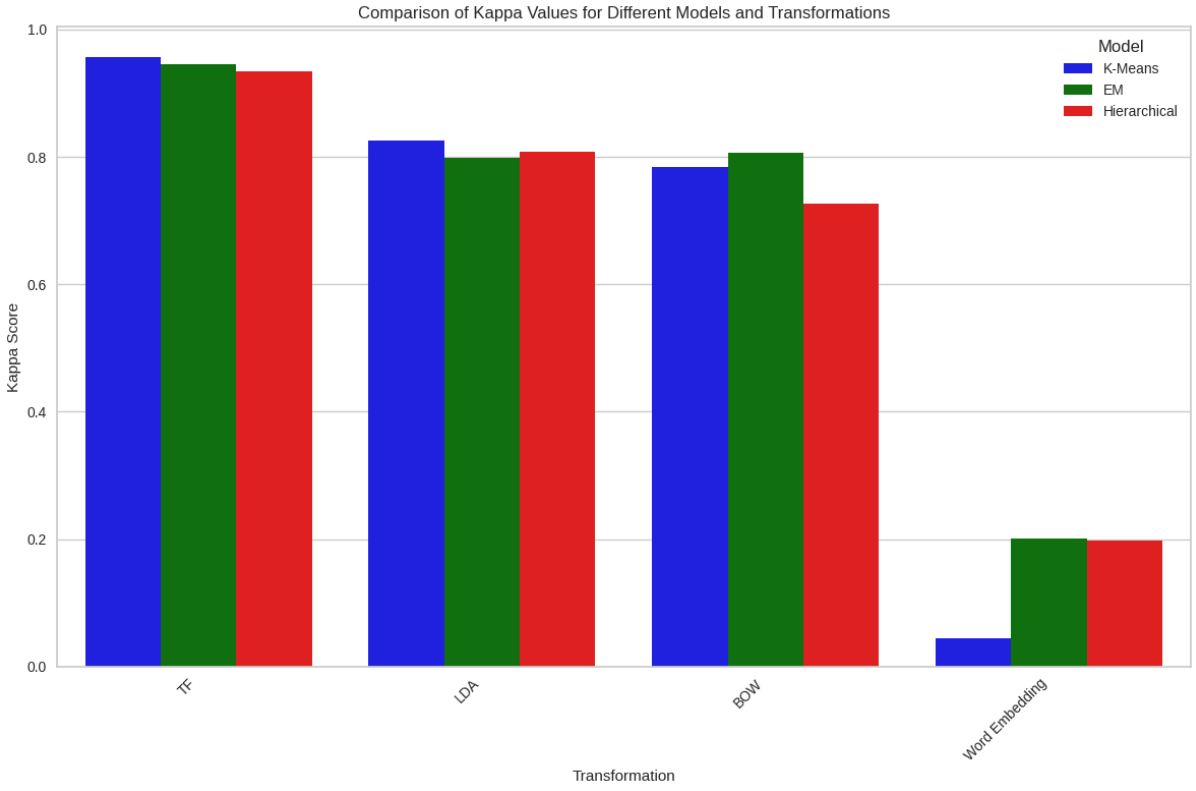

EM Scores for each Transformation using Silhouette value
Score
0.0
0.1
0.2
0.3
0.4
0.5
TF
Word Embedding
BOW
LDA
Transformation


K-Means Scores for each Transformation using Kappa value
Score
0.0
0.2
0.4
0.6
0.8
1.0
TF
LDA
BOW
Word Embedding
Transformation


EM Scores for each Transformation using Kappa value
Score
0.0
0.2
0.4
0.6
0.8
TF
BOW
LDA
Word Embedding
Transformation


Hierarchical Scores for each Transformation using Kappa value
Score
0.0
0.2
0.4
0.6
0.8
TF
LDA
BOW
Word Embedding
Transformation


Comparison of Silhouette Values for Different Models and Transformations
Model
K-Means
Hierarchical
EM
Silhouette Score
0.0
0.1
0.2
0.3
0.4
0.5
Word Embedding
TF
LDA
BOW
Transformation


Comparison of Kappa Values for Different Models and Transformations
Model
K-Means
EM
Hierarchical
Kappa Score
0.0
0.2
0.4
0.6
0.8
1.0
TF
LDA
BOW
Word Embedding
Transformation

Champion model

Champion Model Kappa (First Element):

	0
Model	K-Means
Transformation	TF
Score	0.95625

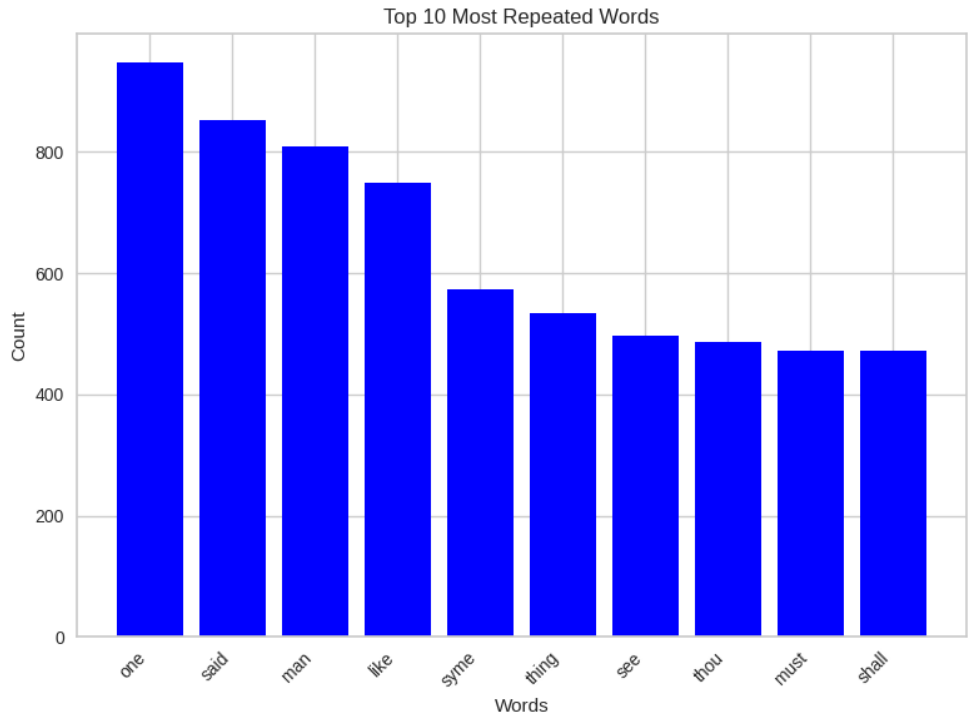
Coherence With LDA

- It's used to measure how well the topics are extracted
- Coherence score With LDA using u-mass method: -2.639600
    - Lower is better
- Coherence score With LDA using c_v method : 0.466800
    - Higher is better

Scores DataFrame:

	Metric	Score
0	Perplexity	-8.410242
1	Coherence (c_v)	0.466800
2	Coherence (u_mass)	-2.639600

Error Analysis



By reducing the number of clusters from 5 to 3

Silhouette DataFrame:

	Model	Transformation	Score
0	EM	Word Embedding	0.519807
1	K-Means	Word Embedding	0.508555
2	K-Means	TF	0.464090
3	EM	TF	0.457543
4	Hierarchical	Word Embedding	0.447223
5	K-Means	BOW	0.442820
6	K-Means	LDA	0.434747
7	Hierarchical	LDA	0.434397
8	Hierarchical	TF	0.428096
9	EM	BOW	0.425220
10	EM	LDA	0.392765
11	Hierarchical	BOW	0.385630

Kappa DataFrame:

	Model	Transformation	Score
0	K-Means	TF	0.594435
1	EM	LDA	0.553571
2	Hierarchical	BOW	0.500153
3	Hierarchical	TF	0.381266
4	EM	BOW	0.317003
5	K-Means	LDA	0.309006
6	Hierarchical	Word Embedding	0.303125
7	EM	Word Embedding	0.246535
8	K-Means	Word Embedding	0.212204
9	Hierarchical	LDA	0.066687
10	EM	TF	0.056839
11	K-Means	BOW	0.016368

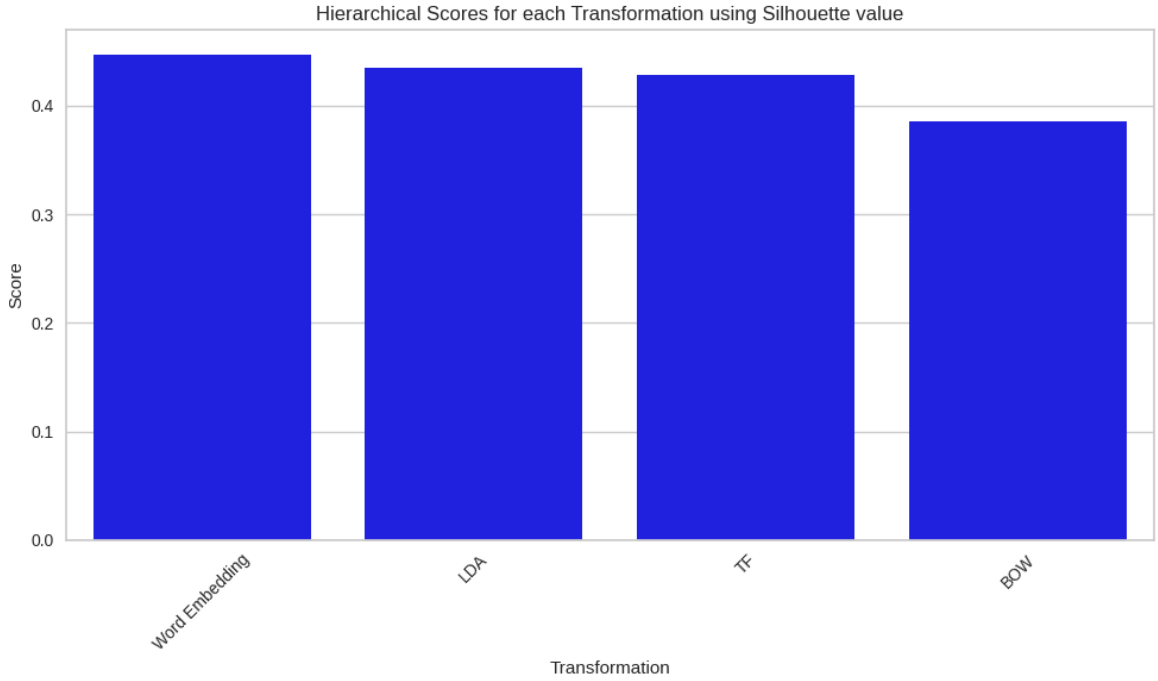
Champion Model Silhouette (First Element):

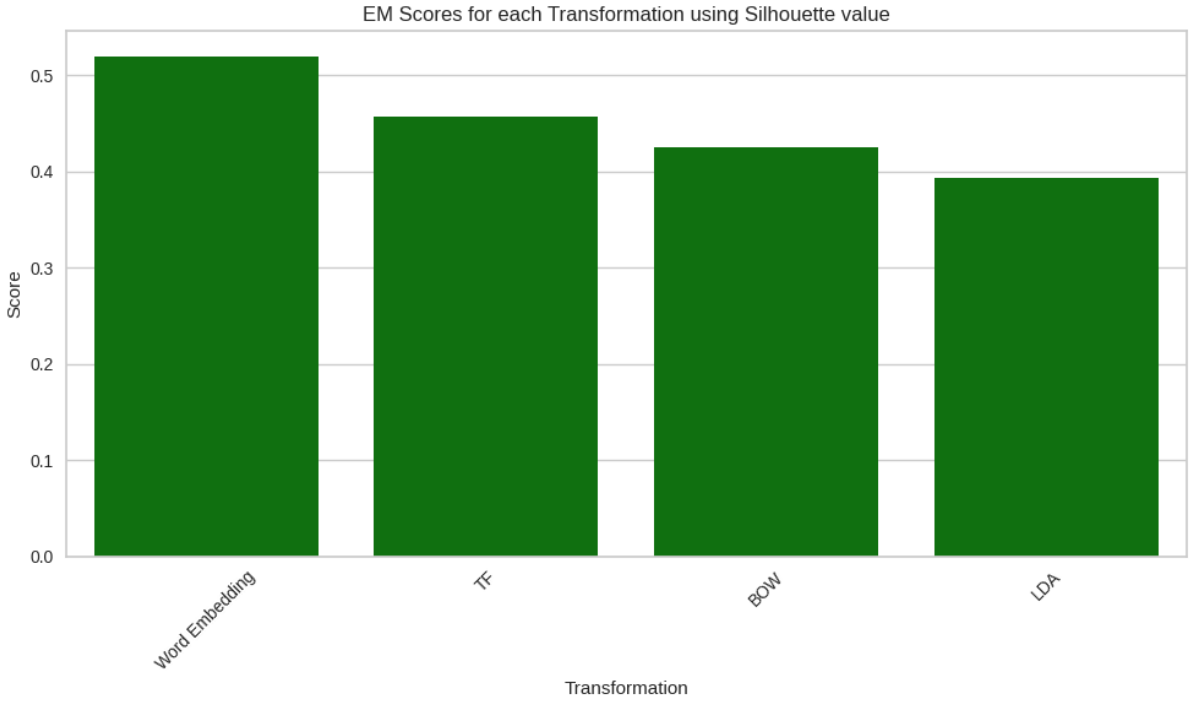
	0
Model	EM
Transformation	Word Embedding
Score	0.519807

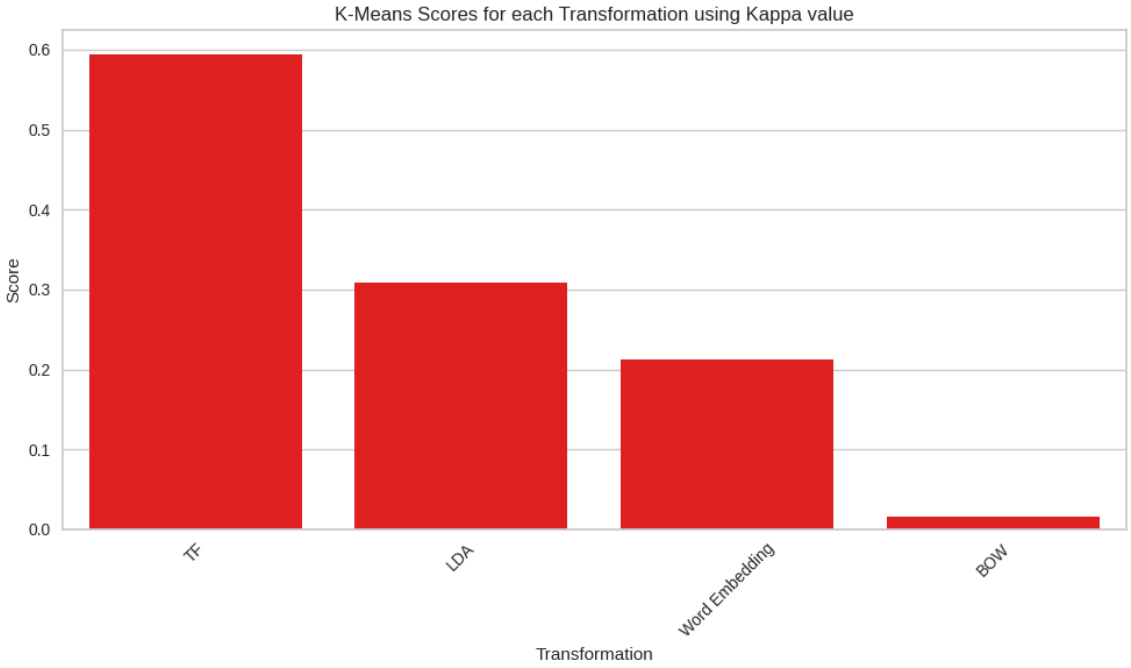
Champion Model Kappa (First Element):

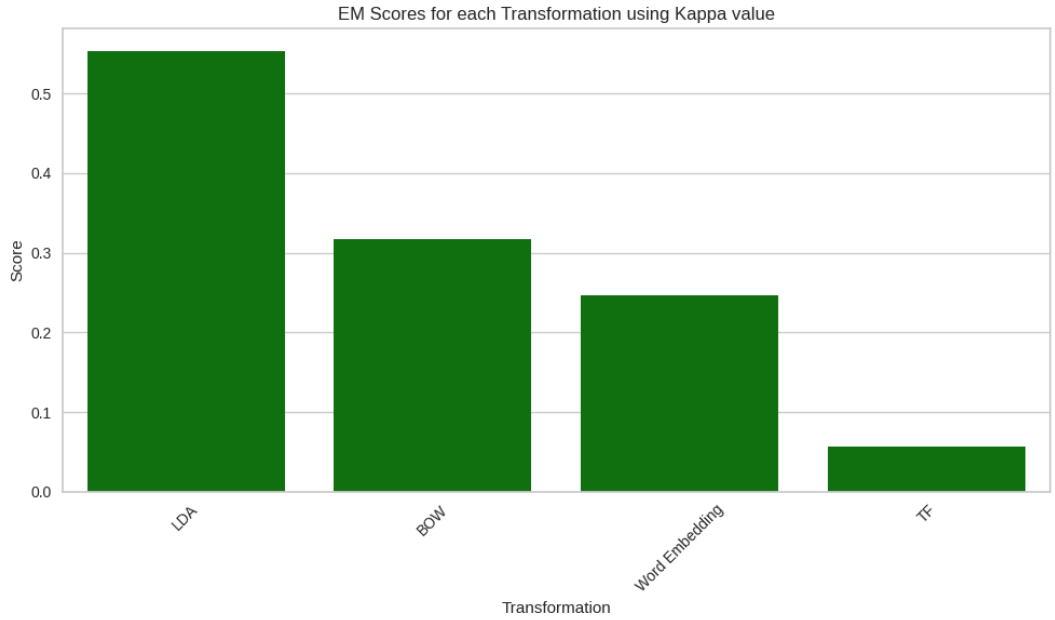
	0
Model	K-Means
Transformation	TF
Score	0.594435

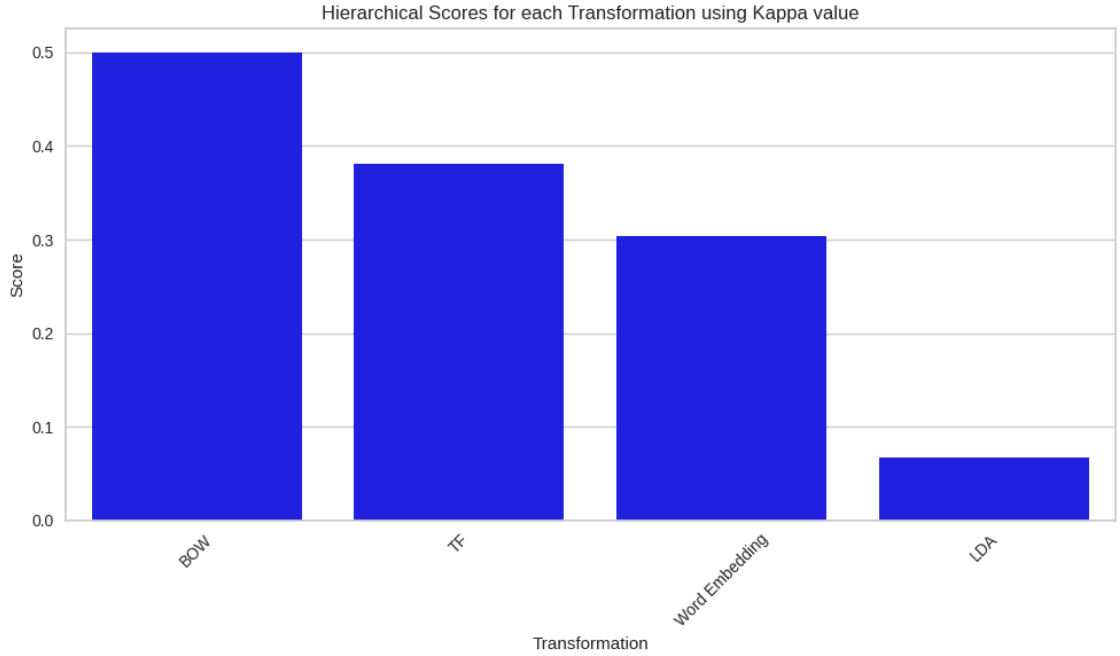

K-Means Scores for each Transformation using Silhouette value
Score
0.0
0.1
0.2
0.3
0.4
0.5
Word Embedding
TF
BOW
LDA
Transformation


Hierarchical Scores for each Transformation using Silhouette value
Score
0.0
0.1
0.2
0.3
0.4
Word Embedding
LDA
TF
BOW
Transformation


EM Scores for each Transformation using Silhouette value
Score
0.0
0.1
0.2
0.3
0.4
0.5
Word Embedding
TF
BOW
LDA
Transformation


K-Means Scores for each Transformation using Kappa value
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
TF
LDA
Word Embedding
BOW
Transformation


EM Scores for each Transformation using Kappa value
Score
0.0
0.1
0.2
0.3
0.4
0.5
LDA
BOW
Word Embedding
TF
Transformation


Hierarchical Scores for each Transformation using Kappa value
Score
0.0
0.1
0.2
0.3
0.4
0.5
BOW
TF
Word Embedding
LDA
Transformation

The best clustering performance was achieved using K-Means with TF representation, while the worst performance was observed with K-Means using Word Embedding. Here are some possible reasons:

1. TF Representation (Best Performance):
  - The term frequency (TF) representation captures the importance of words in documents. This representation likely helps to differentiate the books based on the frequency of relevant terms, which may correlate well with genre distinctions.
  - TF representation is straightforward and interpretable, allowing the K-Means algorithm to find meaningful clusters based on word frequencies.
  -Classics and fiction books may have distinct vocabulary and word usage patterns. TF representation helps to capture these differences, allowing the clustering algorithm to identify and group books based on their term frequencies.

2. Word Embedding (Worst Performance):
  - Word embeddings capture semantic relationships between words. However, in this case, the clustering using Word Embedding with K-Means did not perform well.
  - The poor performance may be due to several reasons:
    - Book data, particularly classics and fiction, might not exhibit strong semantic patterns that can be effectively captured by word embeddings.
    - The quality and coverage of the word embedding model used might not be suitable for the book dataset.
    - Word embeddings can struggle with rare or out-of-vocabulary words, which are prevalent in classic literature.

Comparing K-Means with BOW and K-Means with LDA, both achieved relatively similar results, with slightly higher scores for K-Means with LDA. This suggests that LDA captures underlying topics/themes that help differentiate the books, contributing to better clustering performance.

When analyzing Hierarchical clustering, it generally performed slightly worse than K-Means in terms of silhouette scores. However, it is important to note that the performance of clustering algorithms can vary depending on the dataset and specific characteristics of the books.