DATA SCIENCE APPLICATIONS

Assignment 2

Group members:

Andrew Adel Labib.

Hussien Amin Abdelhafez.

Phoebe Thabit Wadea.

Sandy Adel Latef.

Abstract :

Text clustering, also known as document clustering or text categorization, is a process of grouping similar documents together based on their textual content. It is a common technique used in natural language processing (NLP) and information retrieval to organize large collections of text documents and identify patterns or themes within them.

Introduction

Turning written material into useful information and insights is challenging because text is unstructured and not organized in a standardized way. The general goal is to produce categorizations and predictions from the text, compare the results, evaluate the benefits and limitations of different approaches. To achieve this, we will need to implement the approaches, determine the accuracy of each model, and select the most successful one. Python, with its many libraries, is well suited as a programming language for tackling such problems due to its text processing capabilities.

In our task, we perform these modern clustering techniques on some famous science fiction books which made us easily determine the style of each author later from his future writings thus leading to a big revolution in literature field.

Dataset

The Gutenberg dataset represents a corpus of over 60,000 book texts, their authors and titles. The data has been scraped from the Project Gutenberg website using a custom script to parse all bookshelves. we have taken five different samples of Gutenberg digital books that are of five different authors, that we think are of fiction novels. **The books are**:

- Hot Planet by Hal Clement.
- Foundling on Venus by Dorothy De Courcy and John De Courcy.
- The master mind of Mars by Edgar Rice Burroughs.
- The Red Hell of Jupiter by Paul Ernst.
- Ice Planet by Carl Selwyn.

Data Preparation:

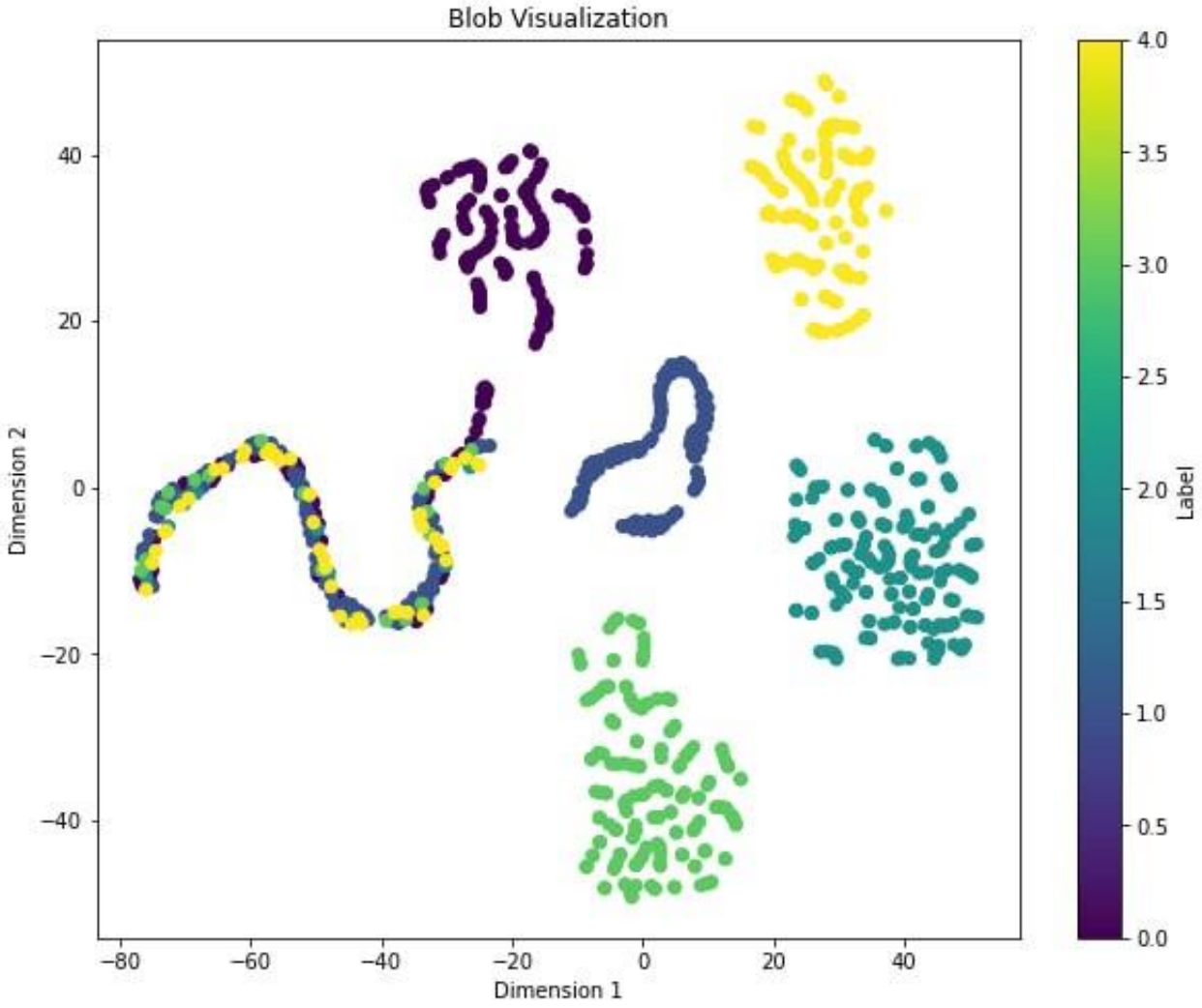
After reading the books, we did some preprocessing methods:

- **Text cleaning** – converting text into lowercase and cleaning it by removing unsubstantial parts such as HTML tags, symbols, or sometimes numbers, ensures that words with less than 3 characters like "ye" are removed as they have no special meaning (Removing non-alphabetic characters)
- **Stop words removal** – excluding some common words that don't provide useful information.
- **Lemmatization and stemming** – simplifying words to their base or root form by following some rules from dictionaries, cutting off common prefixes and suffixes, and similar.
- **Tokenization** – when we separate cleaned text into smaller units, such as words, characters, or some combinations of them.
- **Data Partitioning** – partition each book into 200 documents, each document is a 150-word record.

	partitions	name	author	Label	index
0	agree abide term agreement must cease using re...	Hot Planet	Hal Clement	a	0
1	periodic upheaval heat accumulated inside heat...	Hot Planet	Hal Clement	a	0
2	receiver missing vehicle would detected power ...	Hot Planet	Hal Clement	a	0
3	know speed exactly may two hour maybe five six...	Hot Planet	Hal Clement	a	0
4	constant state change outside united state che...	Hot Planet	Hal Clement	a	0
...	...	...	...	...	...
995	thought came slowly hell inside neptune would ...	Ice Planet	Carl Selwyn	e	4
996	almost restriction whatsoever may copy give aw...	Ice Planet	Carl Selwyn	e	4
997	lipped spun wheel giving enough air slid door ...	Ice Planet	Carl Selwyn	e	4
998	turn take dive like grinned bewildered man wom...	Ice Planet	Carl Selwyn	e	4
999	frozen behind passed place impregnable perfect...	Ice Planet	Carl Selwyn	e	4

1000 rows × 5 columns

Data visualization :



Feature engineering:

- **BOW**: one type of transformations that count of the total occurrences of most frequently used words, to convert the text into numbers so that the algorithm can deal with it.

	abandoned	abandoning	abhorrence	abide	able	ablest	aboard	abrupt	abruptly	absence	...	youthful	zaino	zak	zephyr	zero	zinc	zip	zone	zoom	zoomed
0	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	1	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0	0	0	...	0	1	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
996	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
997	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
998	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
999	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

1000 rows × 7242 columns

- **TF_IDF**: term frequency-inverse document frequency is a measure for estimating the importance of words in a document among a collection of documents.

	abandoned	abandoning	abduction	abide	able	ablest	aboard	abroad	abruptly	absence	...	zaino	zak	zephyr	zero	zinc	zip	zode	zone	zoom	zoomed
0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.000000	0.0	0.070352	0.0	0.0	0.0	...	0.054160	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.068476	0.0	0.072988	0.0	0.0	0.0	...	0.112378	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
996	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
997	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
998	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
999	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

1000 rows × 7278 columns

- **LDA :** is a generative probabilistic model used for topic modeling. It assumes that each document in a corpus is a mixture of various topics,

and each topic is a distribution over words. LDA represents documents as a distribution over topics, and words as a distribution over topics.

```
Average topic coherence: -2.1635.
[([(0.06660278, 'project'),
   (0.060370512, 'work'),
   (0.021994635, 'gutenberg'),
   (0.020473476, 'electronic'),
   (0.017090308, 'term'),
   (0.016377488, 'state'),
   (0.015301475, 'foundation'),
   (0.014332525, 'copy'),
   (0.013147522, 'copyright'),
   (0.012830056, 'donation'),
   (0.012335455, 'license'),
   (0.011794849, 'full'),
   (0.011284476, 'agreement'),
   (0.011013283, 'literary'),
   (0.010889794, 'archive'),
   (0.010812877, 'paragraph'),
   (0.010226998, 'united'),
   (0.010204196, 'fee'),
   (0.009511911, 'may'),
   (0.009482837, 'use')],
```

- **Word Embedding:** refers to the representation of words as dense vectors in a high-dimensional space, where each dimension represents a unique aspect or feature of the word's meaning. Word embeddings are used to capture semantic and syntactic relationships between words.

```
array([ 0.02893017,  0.17034206, -0.10192718, -0.02860261, -0.0685344 ,
       -0.02527214,  0.081892  ,  0.35390922, -0.3855857 ,  0.00514269,
       -0.17995624, -0.12912057,  0.17674202, -0.11147236, -0.0785943 ,
       -0.08926583, -0.02395255,  0.00392741,  0.14649384, -0.19034879,
       -0.08966579, -0.18568155,  0.0254396 ,  0.00416615, -0.11773852,
        0.12091141, -0.11304222, -0.22244988, -0.09362275, -0.07000446,
       -0.01908687,  0.10792698, -0.14051856,  0.05430949, -0.0161921 ,
        0.07119137, -0.17170045, -0.43667397, -0.09776276, -0.43426606,
        0.07114695, -0.19212937, -0.09609844, -0.03677533,  0.05512959,
       -0.18438408,  0.16714351, -0.09738971, -0.02938952, -0.16281185,
       -0.05384807, -0.03259765, -0.2641178 , -0.02336095, -0.16929844,
       -0.06245825, -0.1709215 , -0.05281397, -0.17780152, -0.03065702,
        0.21143611,  0.08818959, -0.23649392, -0.10151506, -0.3162379 ,
        0.47919884,  0.06184378,  0.19996041, -0.33729553,  0.06087659,
       -0.27754354,  0.20263664,  0.09610207, -0.03276955,  0.39127406,
        0.2960457 ,  0.01211386, -0.22652519, -0.05965605,  0.07614165,
        0.02021304, -0.15527922, -0.04569517,  0.1214036 , -0.1717668 ,
        0.2027979 , -0.05706021, -0.0568261 ,  0.27628353, -0.03408328,
       -0.00808412, -0.15657946, -0.06879705,  0.14675285,  0.06946749,
        0.03597596,  0.24203783, -0.03795389, -0.10812977, -0.24869306],
      dtype=float32),
array([-0.01098119,  0.1731292 , -0.01559804,  0.0032303 , -0.00290466,
       -0.138114  ,  0.0571065 ,  0.4891066 , -0.36144704, -0.09548447,
       -0.1468073 , -0.255633  ,  0.10927998, -0.01115642, -0.14735955,
       -0.1132386 ,  0.06801768, -0.02444804,  0.10937588, -0.2637634 ,
       -0.06385977, -0.252251  ,  0.09970562, -0.21033347, -0.05243052,
        0.16206606, -0.14673153, -0.23114708, -0.18358068, -0.17184302,
       -0.00750474,  0.21552703, -0.13648722, -0.04250111, -0.03232206,
        0.04064763, -0.20176299, -0.44284305, -0.27572322, -0.37753454,
        0.06191266, -0.20326613, -0.1278697 , -0.19333188,  0.04206033,
```

Clustering :

For each technique of the above, these following models are trained and tested :

- K_means .
- EM.
- Hierarchical clustering.

- **Kappa Score :**

	K-means	**EM**	**Hierarchical clustering**
BOW	0.4425	0.3087	0.2013
TF-IDF	0.2775	0.294	0.0400
LDA	0.615	0.4975	0.6463
Word Embedding	0.1149	0.1513	0.0275

Mapping function : in Kappa score we need to map the actual labels of the data into the labels that produced from models , which map the partition to the most frequent label , in this partition .

- **Silhouette score:**

	K-means	**EM**	**Hierarchical clustering**
BOW	0.0831	0.0691	0.0765
TF-IDF	0.0699	0.0533	0.0757
LDA	0.514	0.3992	0.5071
Word Embedding	0.1975	0.1664	0.1975

Average topic coherence: -1.24620

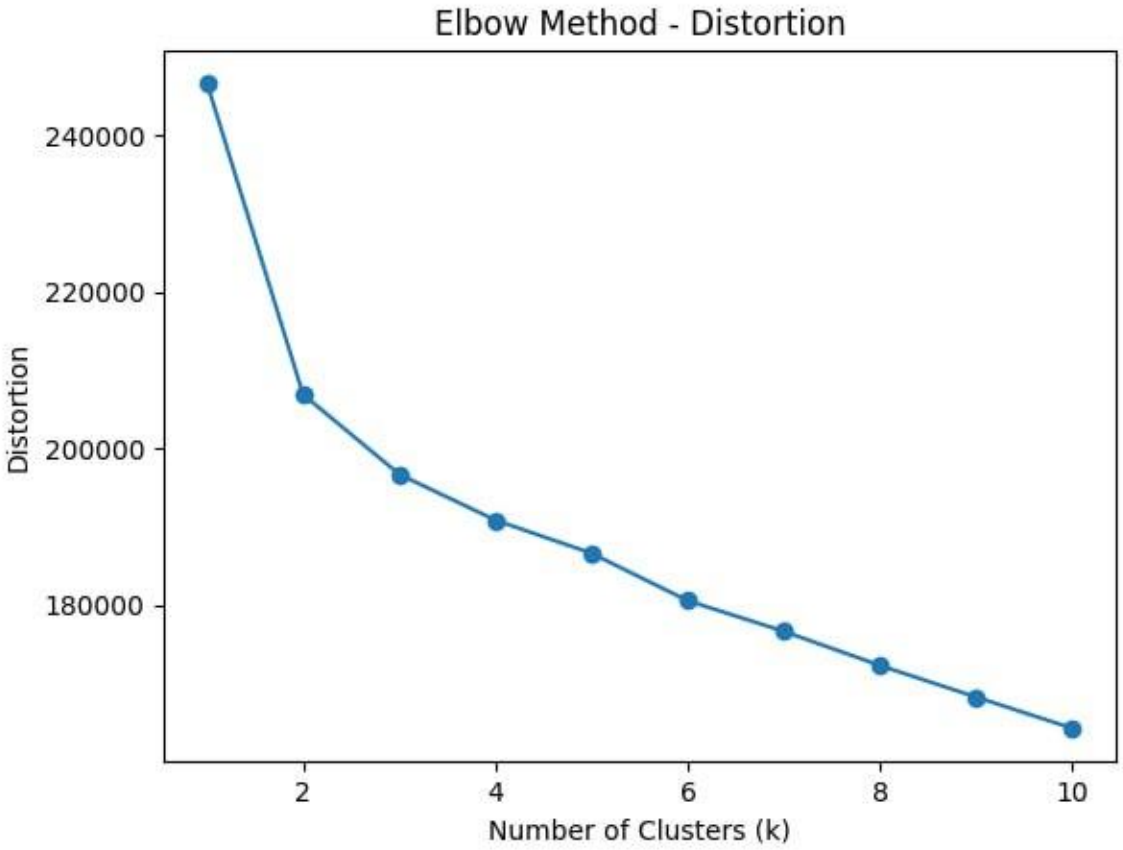
Error Analysis:

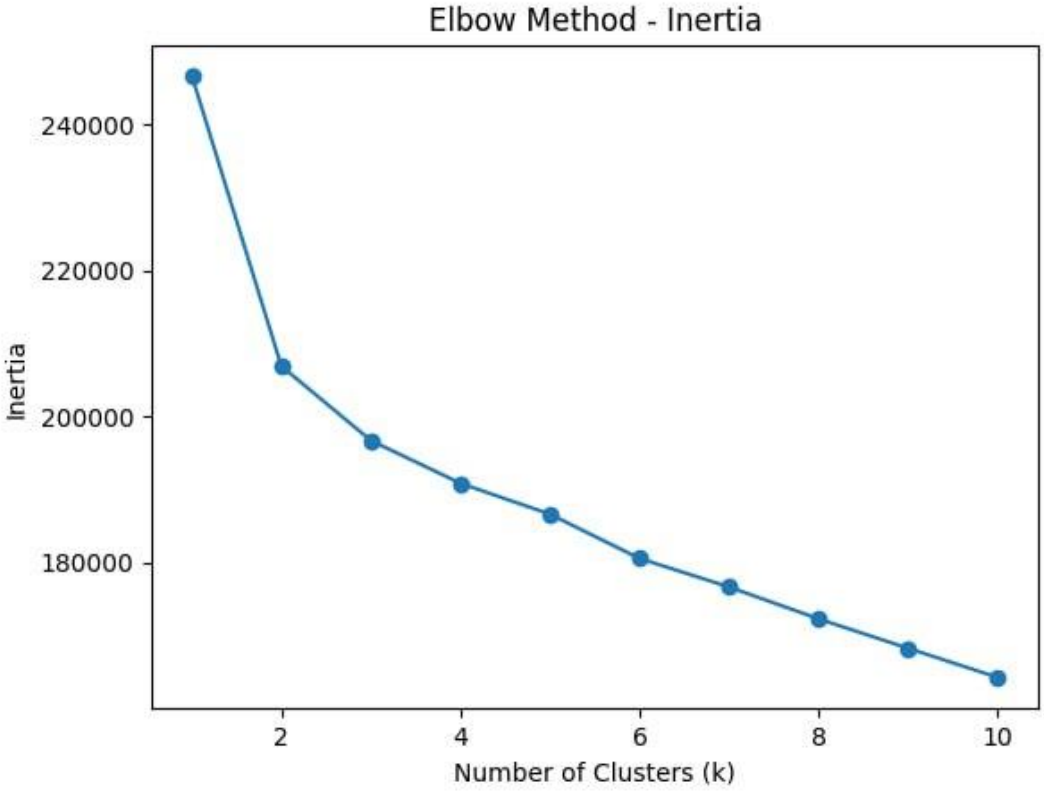
- **Elbow method :** Elbow plot showing the relationship between the number of clusters (k) and the inertia value. Inertia represents the within-cluster sum of squares, which measures the compactness of

clusters. The plot demonstrates that as the number of clusters increases, the inertia decreases, indicating better clustering performance.

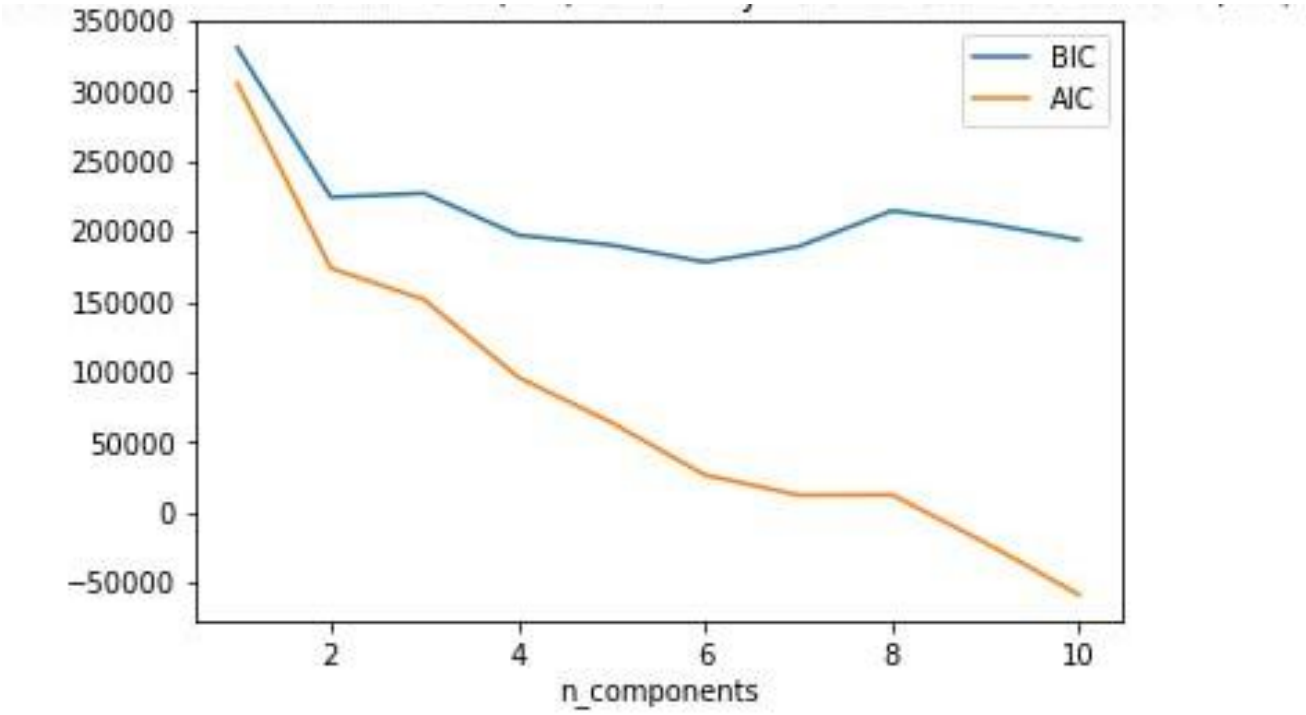
- **Akaike information criterion (AIC) or the Bayesian information criterion (BIC**): is used to compare different models based on their goodness of fit and complexity. The AIC takes into account the likelihood of the model and penalizes it based on the number of parameters used in the model. The goal is to find the model that maximizes the likelihood while minimizing the number of parameters, the BIC balances the goodness of fit and model complexity.

1. **BOW :**
    - **Kmeans on BOW (Elbow Method):**





- **Akaike information criterion (AIC) or the Bayesian information criterion (BIC):**

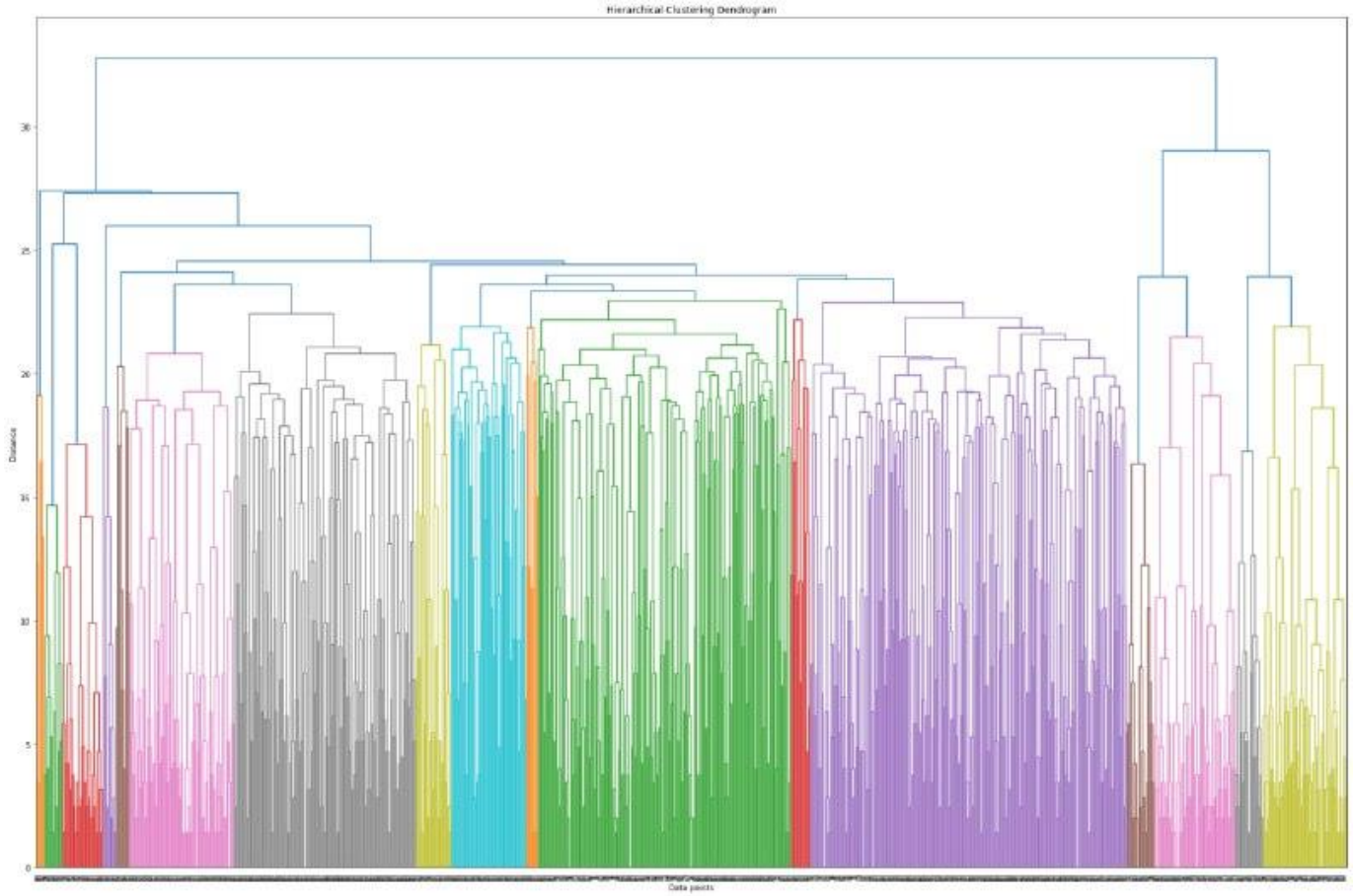


The figure shows the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores for different numbers of components in a Gaussian Mixture Model. The AIC and BIC scores are commonly

used for model selection and can help determine the optimal number of components for the model.

In this plot, we can observe that both the AIC and BIC scores decrease as the number of components increases.

- **Hierarical Clustering :**



Error Analysis (Using adjusted Rand index)

```
from sklearn.metrics import adjusted_rand_score
from sklearn.metrics import fowlkes_mallows_score

# Calculate adjusted Rand index
ari_score = adjusted_rand_score(human_label, hierarchical_labels)

print("Adjusted Rand index:", ari_score)
```

```
Adjusted Rand index: 0.1954633615937071
```

Error Analysis (Using Fowlkes-Mallows index)

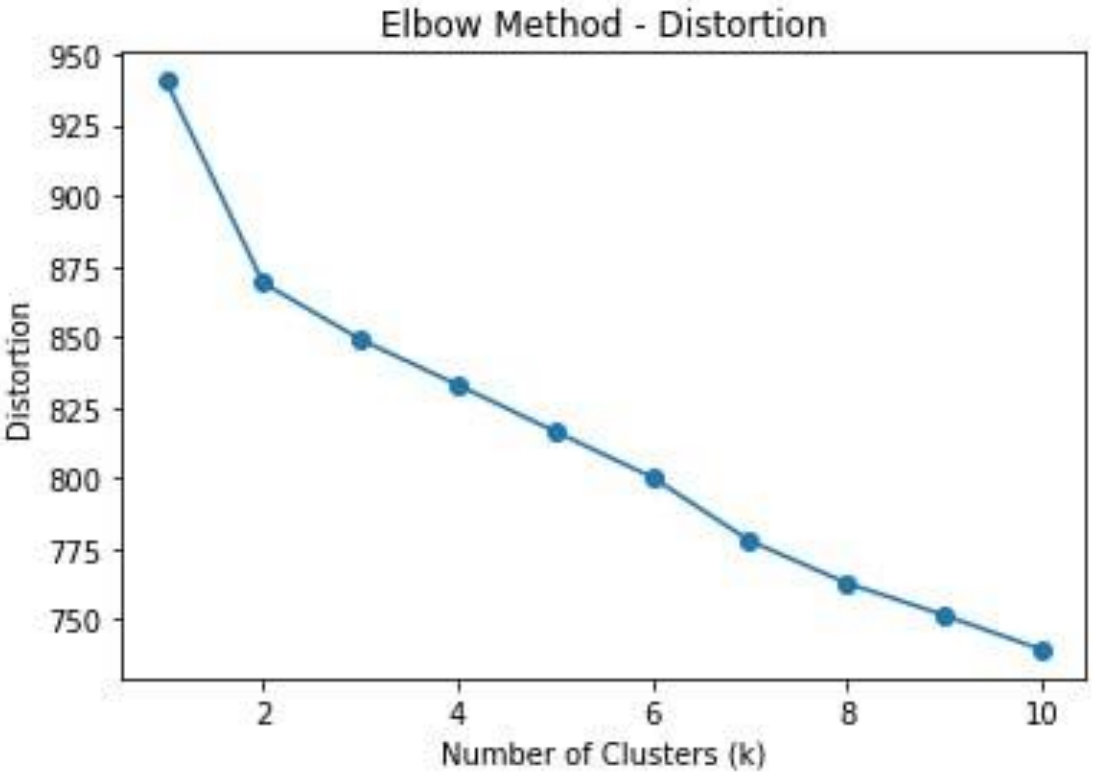
```
# Calculate Fowlkes-Mallows index
fm_score = fowlkes_mallows_score(human_label, hierarchical_labels)

print("Fowlkes-Mallows index:", fm_score)
```

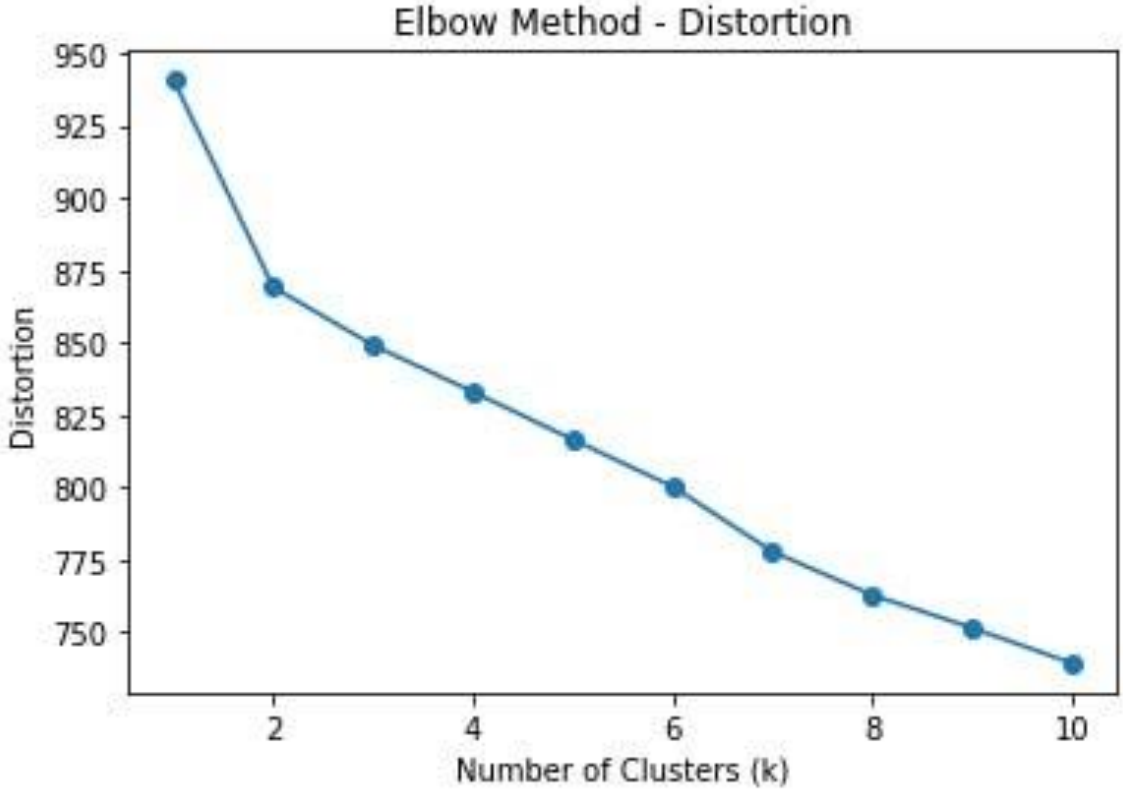
```
Fowlkes-Mallows index: 0.4454864634137153
```

The Adjusted Rand Index measures the similarity between the clustering results and the true labels, taking into account all pairs of samples and their assignments. A higher ARI score indicates a better clustering result, with a score of 1 indicating a perfect match between the clustering and the ground truth.

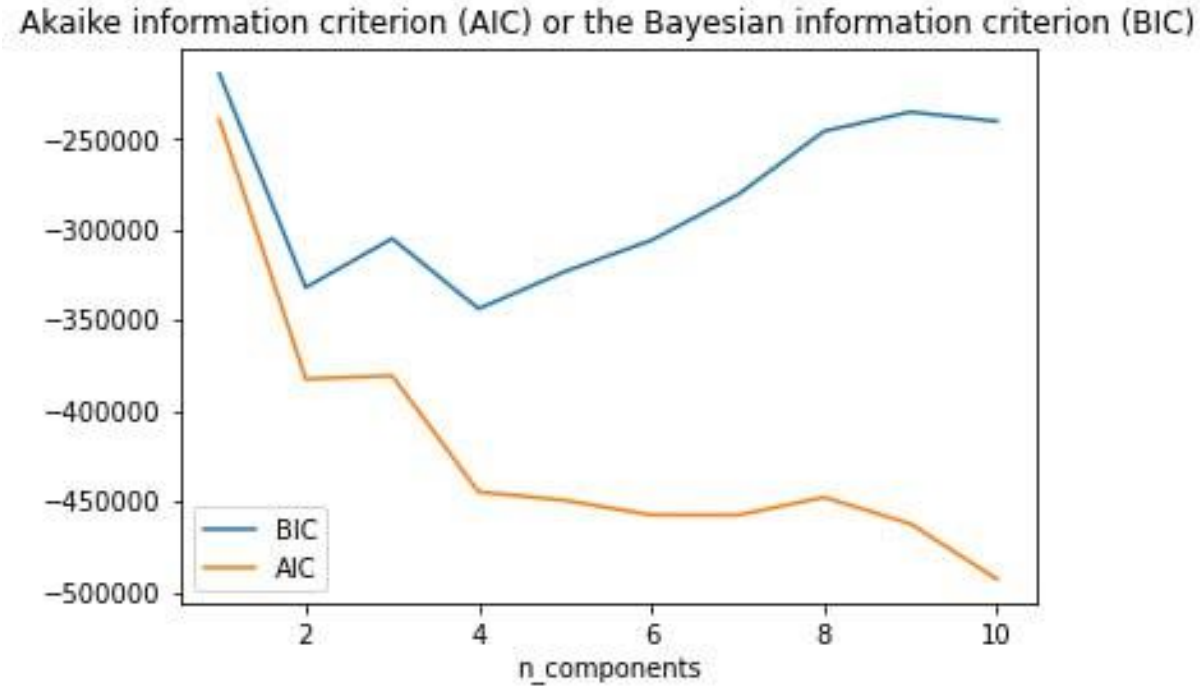
The Fowlkes-Mallows Index evaluates the similarity between the clustering and the true labels based on the pairwise similarities between samples. It considers the number of pairs that are correctly assigned within clusters and the number of pairs that are correctly assigned across different clusters. A higher FMI score indicates a better clustering result, with a score of 1 indicating a perfect match.

2. **TF-IDF:**
    a. **TF-IDF on BOW (Elbow Method):**

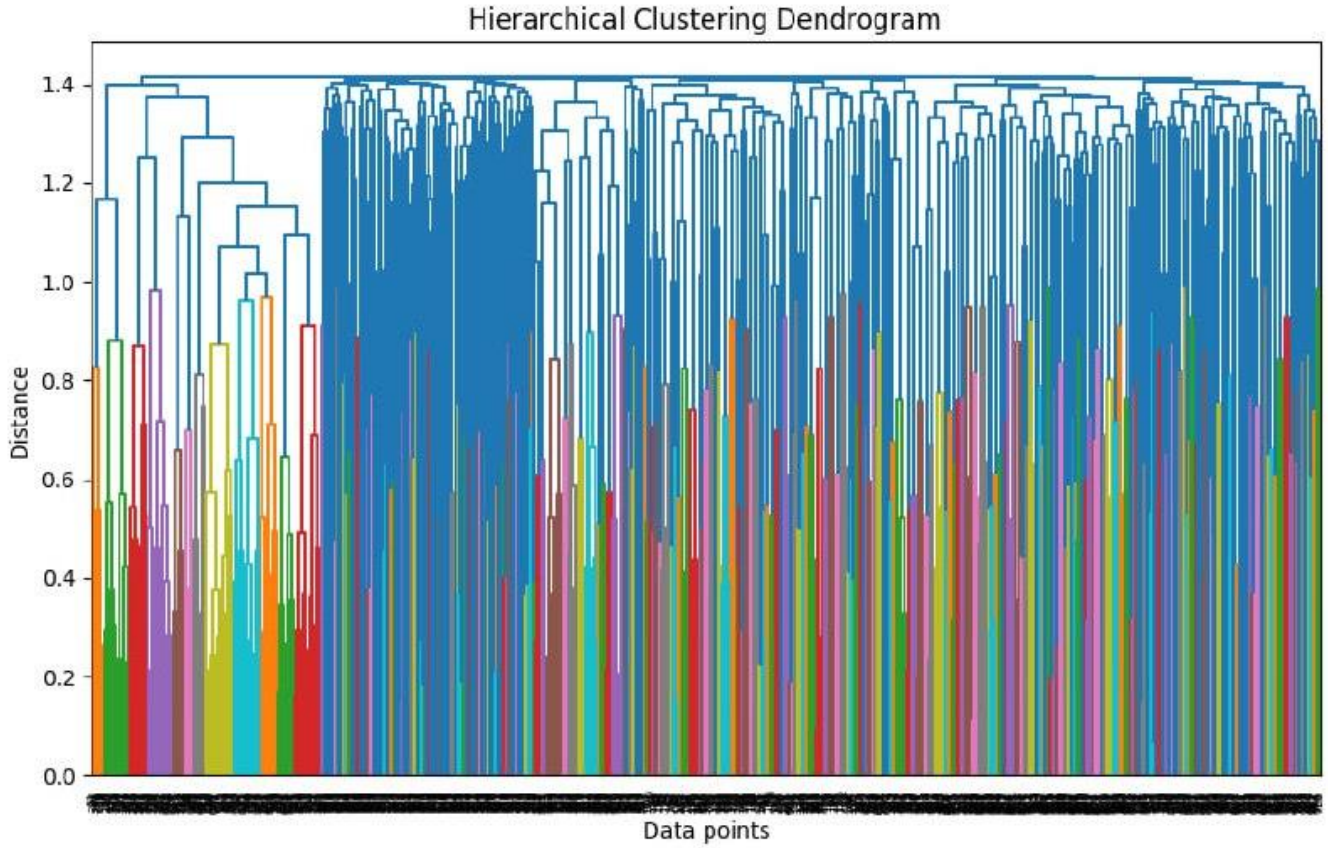

Elbow Method - Distortion
Distortion
950
925
900
875
850
825
800
775
750
2
4
6
8
10
Number of Clusters (k)


Elbow Method - Distortion
Distortion
950
925
900
875
850
825
800
775
750
2
4
6
8
10
Number of Clusters (k)

b. AIC and BIC :



c. Hierarchical Clustering:



Error Analysis (Using adjusted Rand index)

```
from sklearn.metrics import adjusted_rand_score
from sklearn.metrics import fowlkes_mallows_score

# Calculate adjusted Rand index
ari_score = adjusted_rand_score(human_label, hierarchical_labels)

print("Adjusted Rand index:", ari_score)
```

```
Adjusted Rand index: 0.10215195192974924
```

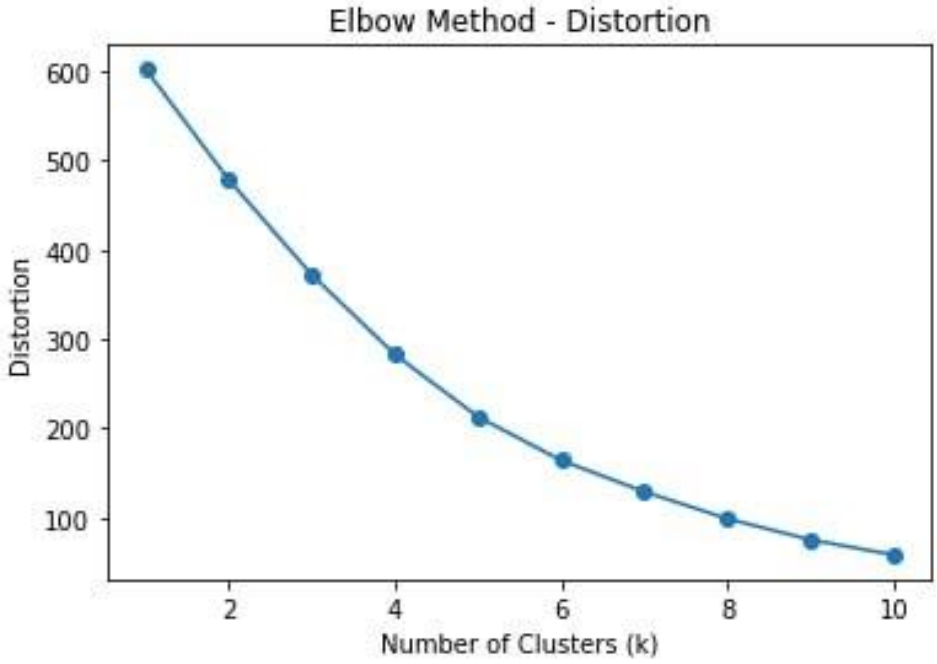
Error Analysis (Using Fowlkes-Mallows index)

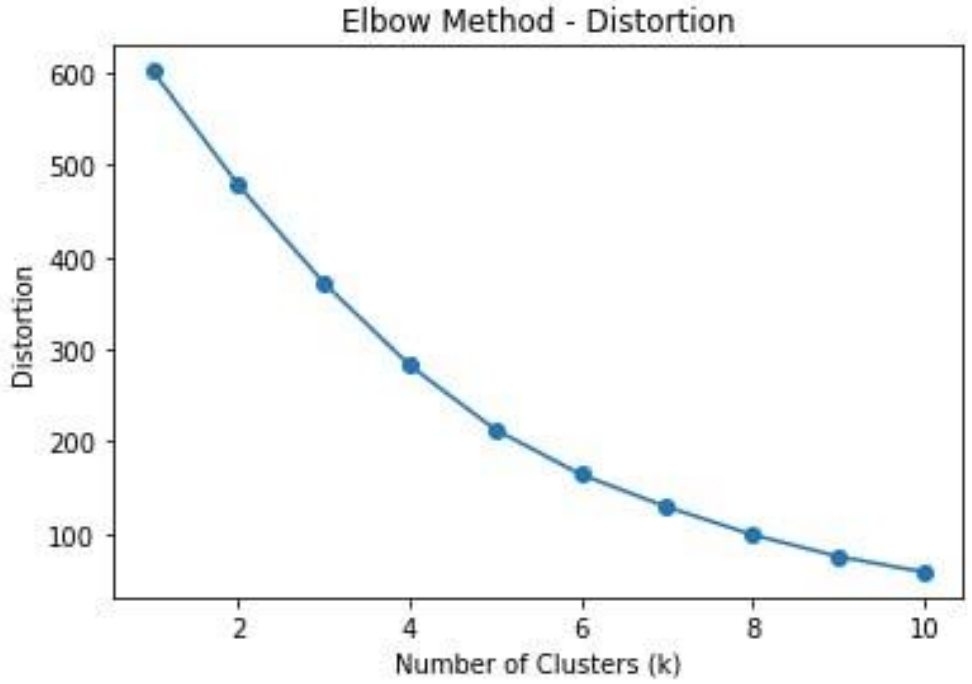
```
# Calculate Fowlkes-Mallows index
fm_score = fowlkes_mallows_score(human_label, hierarchical_labels)

print("Fowlkes-Mallows index:", fm_score)
```

```
Fowlkes-Mallows index: 0.400255199254926
```

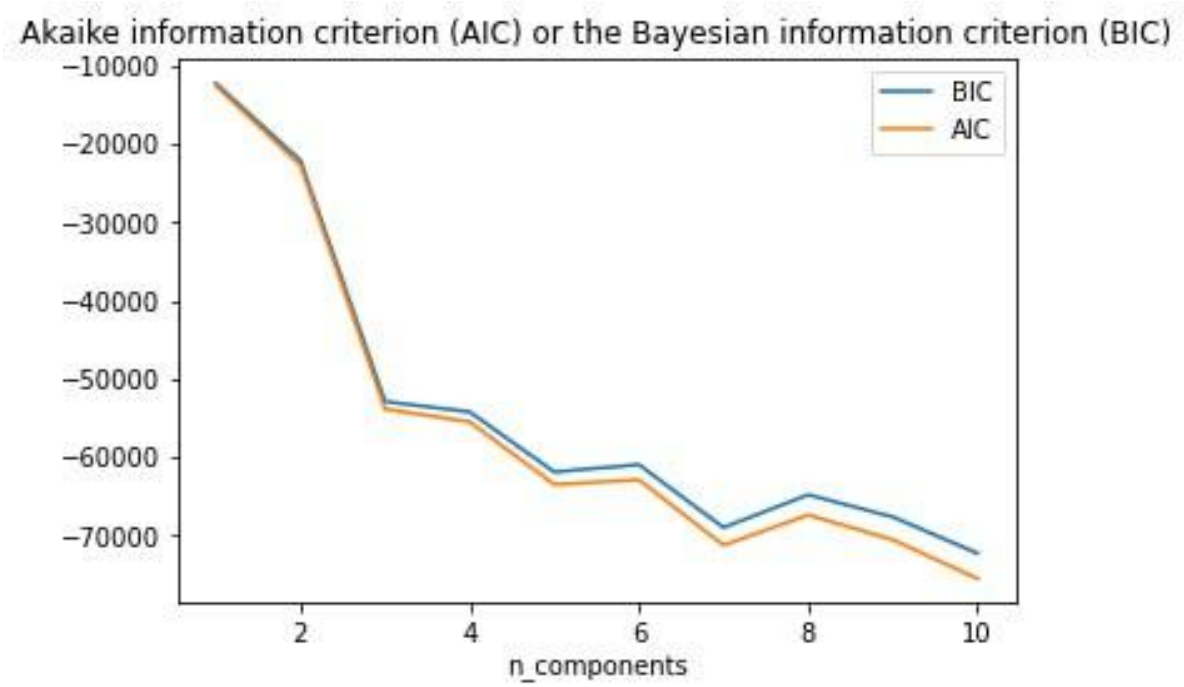
3. LDA:

- **Elbow methos :**




Elbow Method - Distortion
Distortion
600
500
400
300
200
100
2
4
6
8
10
Number of Clusters (k)

- **AIC and BIC:**


Akaike information criterion (AIC) or the Bayesian information criterion (BIC)
−10000
−20000
−30000
−40000
−50000
−60000
−70000
BIC
AIC
2
4
6
8
10
n_components

Error Analysis (Using adjusted Rand index)

```
from sklearn.metrics import adjusted_rand_score
from sklearn.metrics import fowlkes_mallows_score

# Calculate adjusted Rand index
ari_score = adjusted_rand_score(human_label, hierarchical_labels)

print("Adjusted Rand index:", ari_score)
```

```
Adjusted Rand index: 0.5124380730649554
```

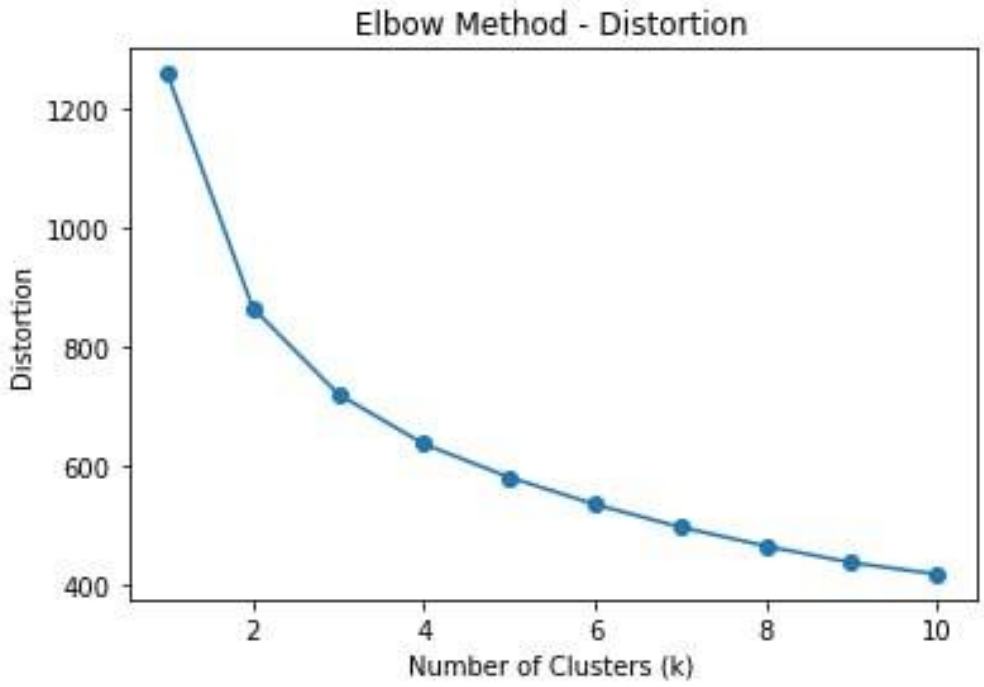
Error Analysis (Using Fowlkes-Mallows index)

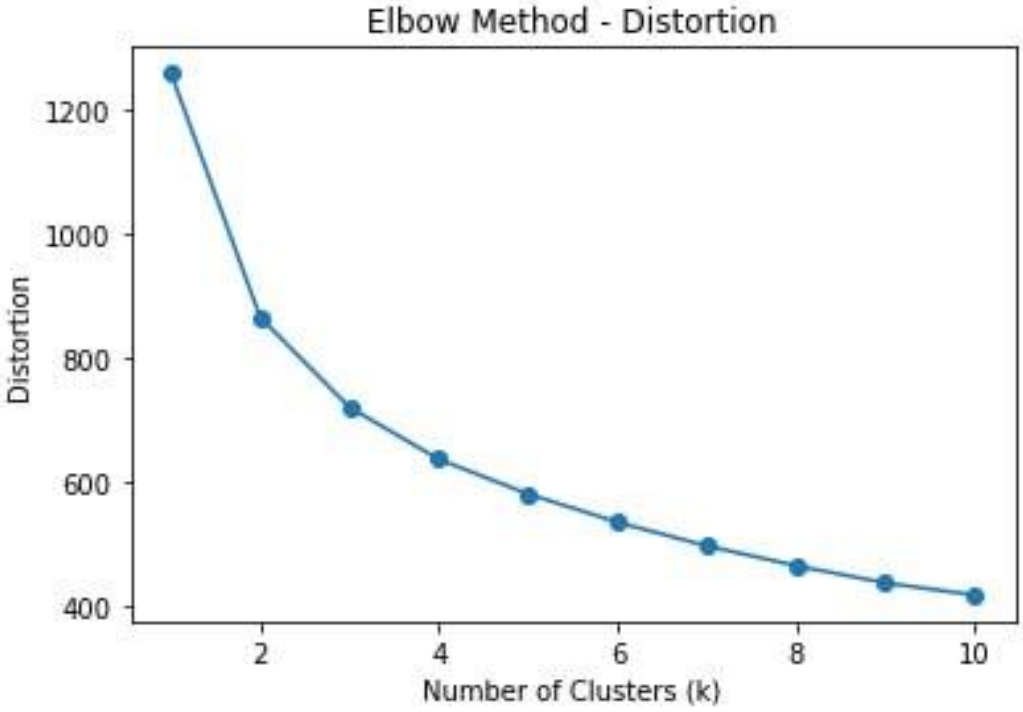
```
# Calculate Fowlkes-Mallows index
fm_score = fowlkes_mallows_score(human_label, hierarchical_labels)

print("Fowlkes-Mallows index:", fm_score)
```

```
Fowlkes-Mallows index: 0.6165768009879496
```

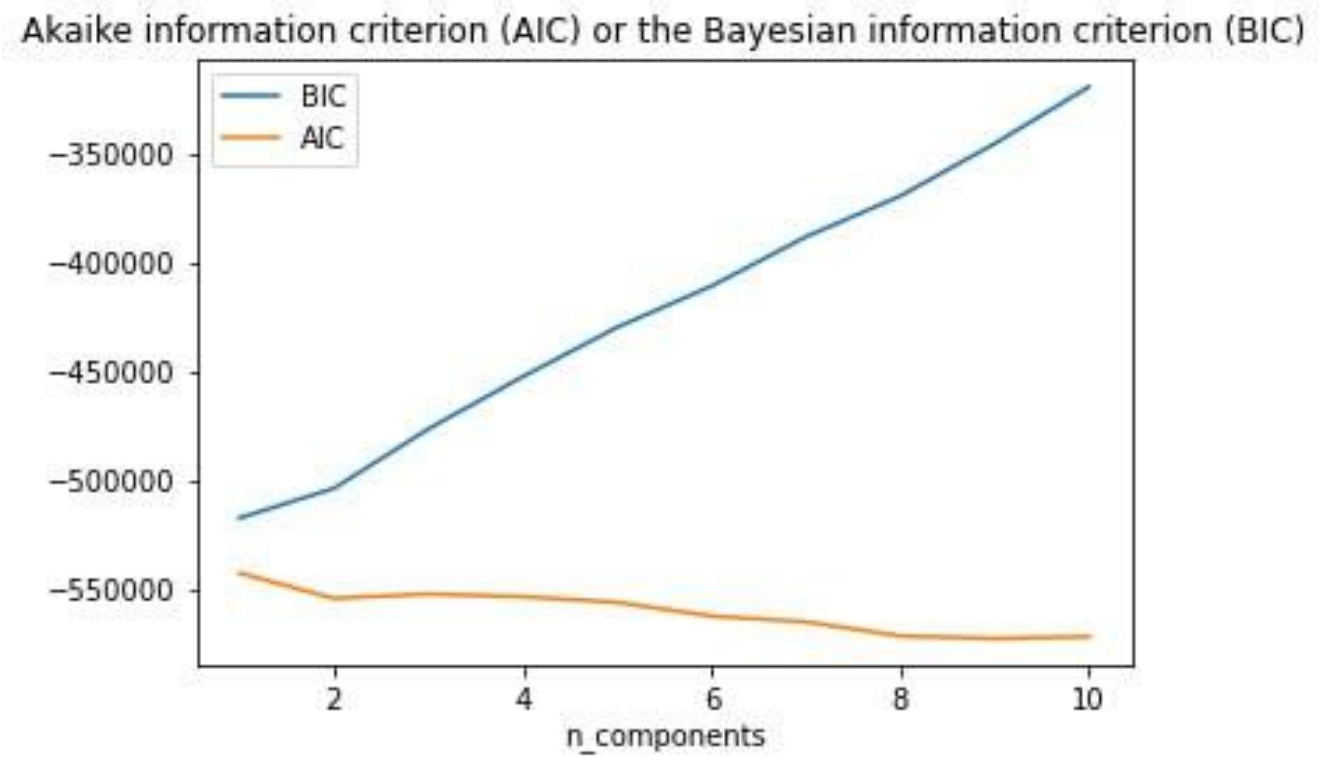
4.Word Embedding:

- **Elbow method:**

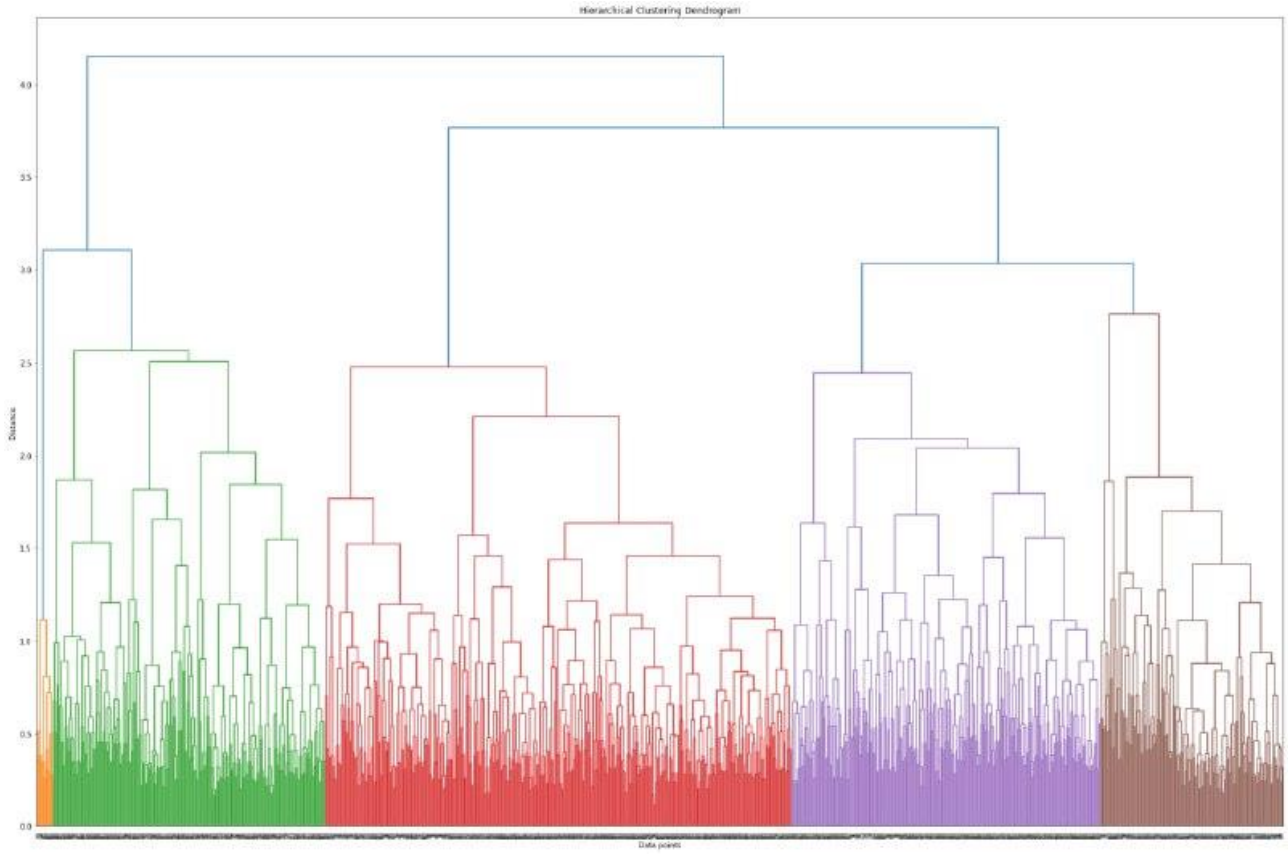




- **AIC and BIC:**



- **Hierarchical Clustering :**



Error Analysis (Using adjusted Rand index)

```
from sklearn.metrics import adjusted_rand_score
from sklearn.metrics import fowlkes_mallows_score

# Calculate adjusted Rand index
ari_score = adjusted_rand_score(human_label, hierarchical_labels)

print("Adjusted Rand index:", ari_score)
```

```
Adjusted Rand index: 0.07715360893069159
```

Error Analysis (Using Fowlkes-Mallows index)

```
# Calculate Fowlkes-Mallows index
fm_score = fowlkes_mallows_score(human_label, hierarchical_labels)

print("Fowlkes-Mallows index:", fm_score)
```

```
Fowlkes-Mallows index: 0.2906102486275637
```

Champion model :

Our champion model is : Hierarchical Clustering in LDA .

After changing parameters in our champion model , its performance become worse .