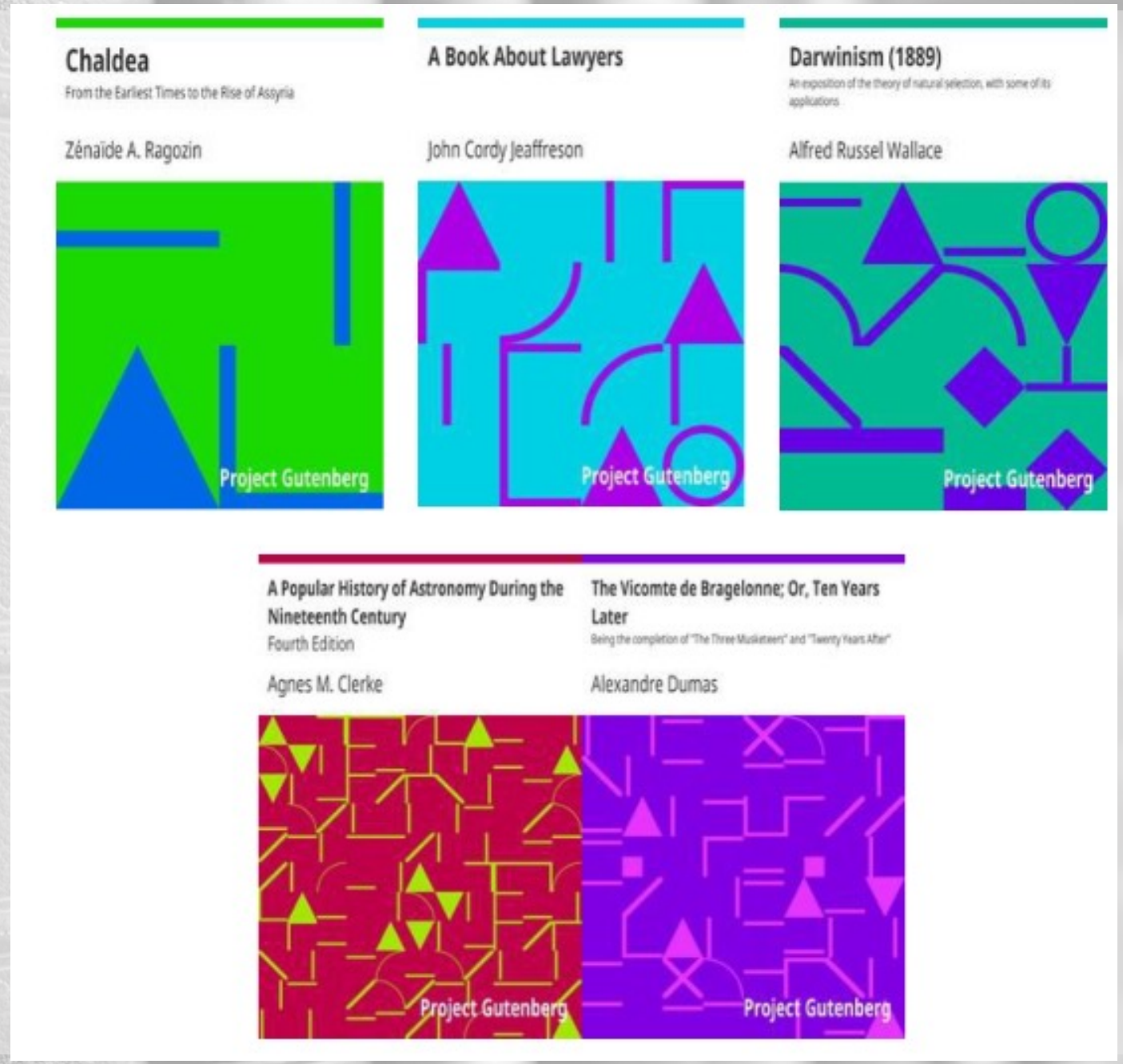
Text clustering model

Himani Kaushik

Need:

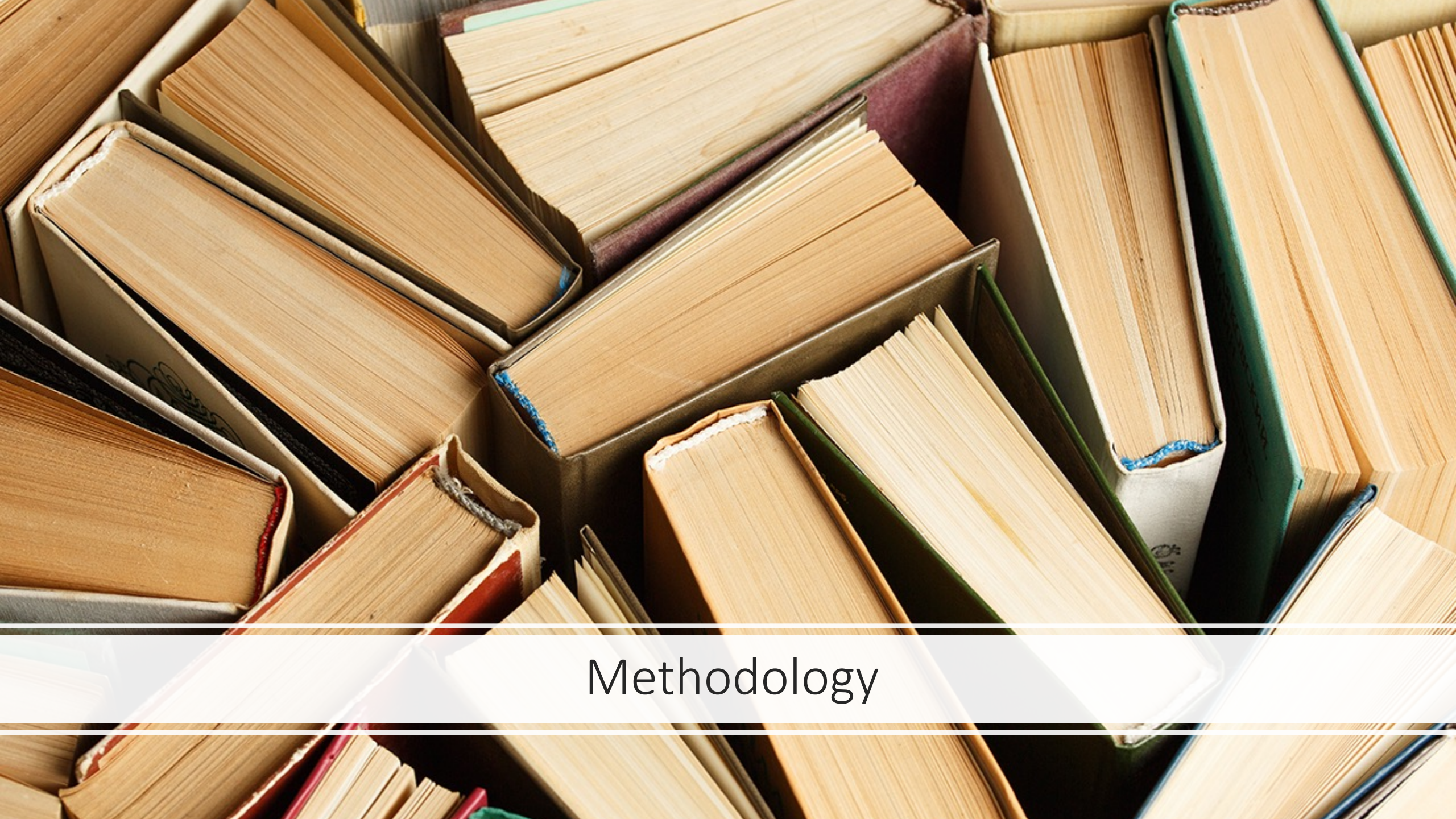
- Categorizing system based on similarity:
  - Authors
  - Genres
- Text transformation and clustering techniques
- Best algorithm to assign correct labels

Data: Project Gutenberg

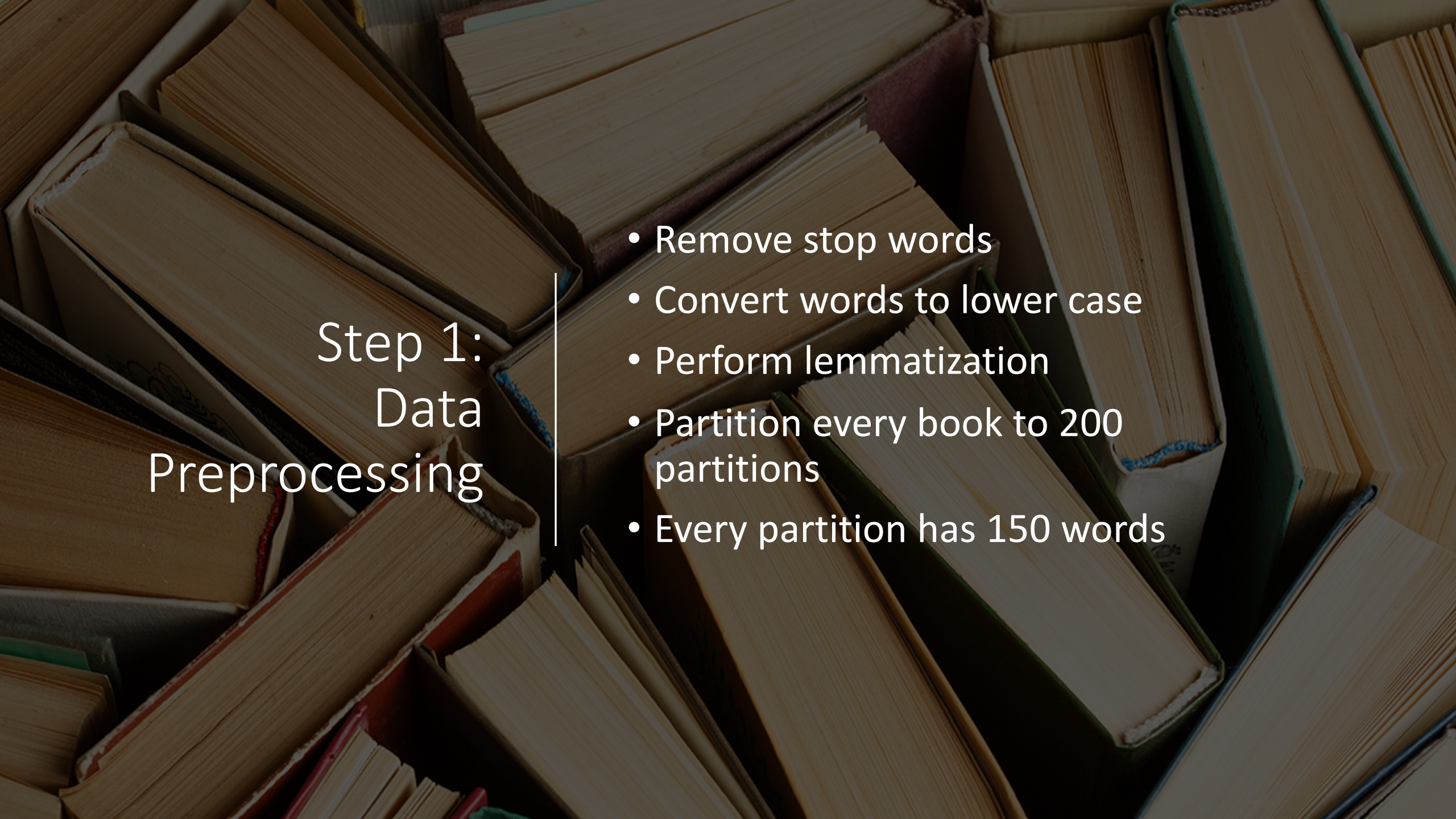


Tools:

- Requests
- Pandas and Numpy
- NLTK, gensim, scikit-learn
- Matplotlib, Seaborn, Wordcloud, pyLDAvis, scipy



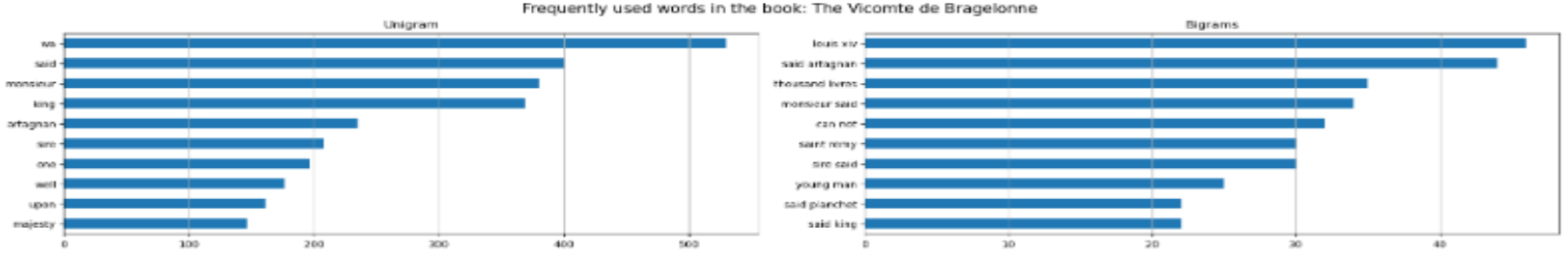
Methodology

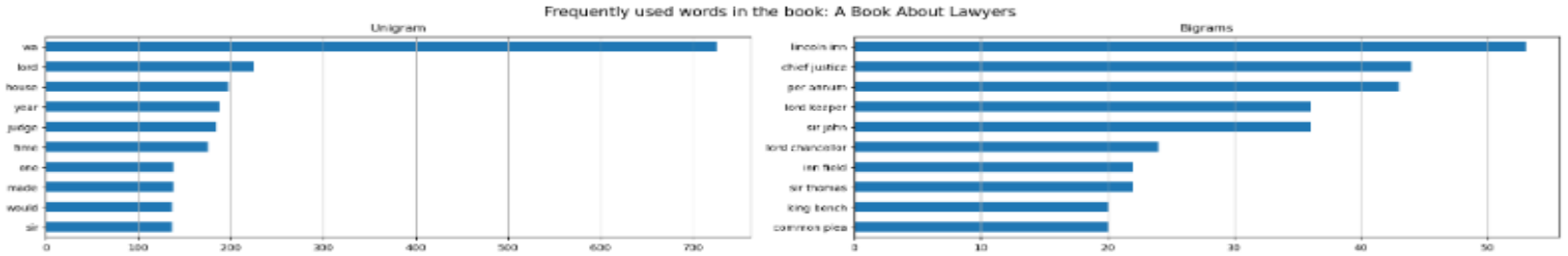


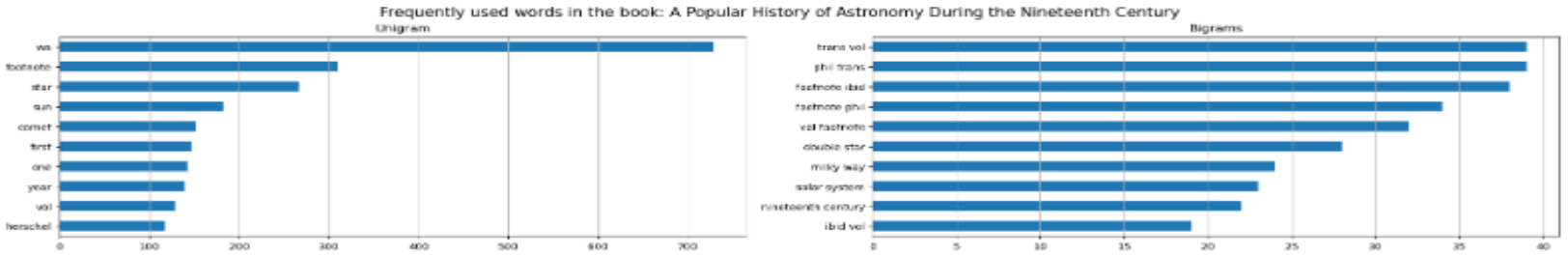
Step 1: Data Preprocessing

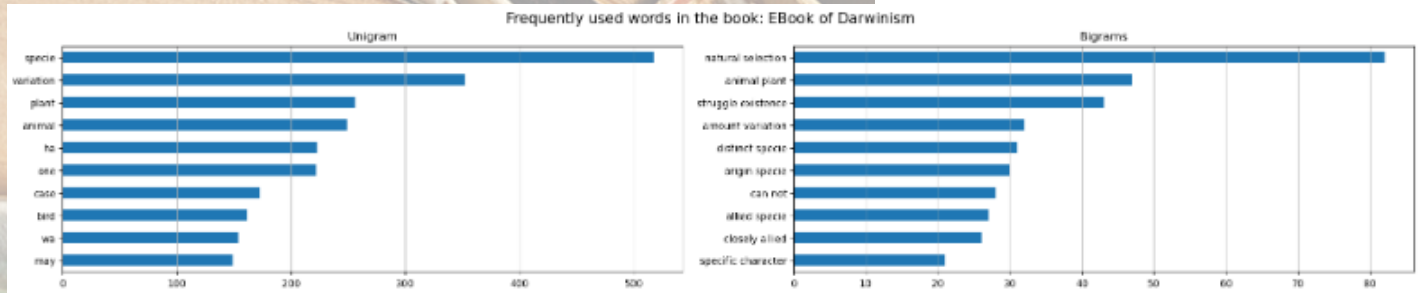
- Remove stop words
- Convert words to lower case
- Perform lemmatization
- Partition every book to 200 partitions
- Every partition has 150 words

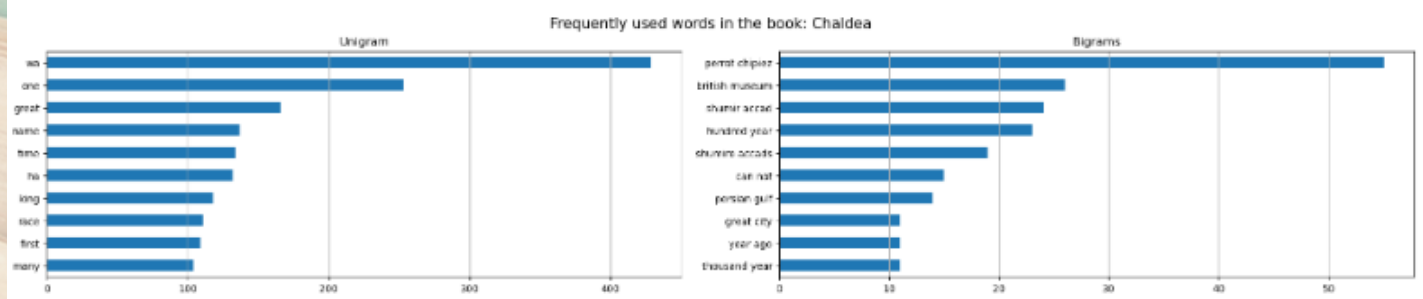
Frequent Words: Unigrams and Bigrams





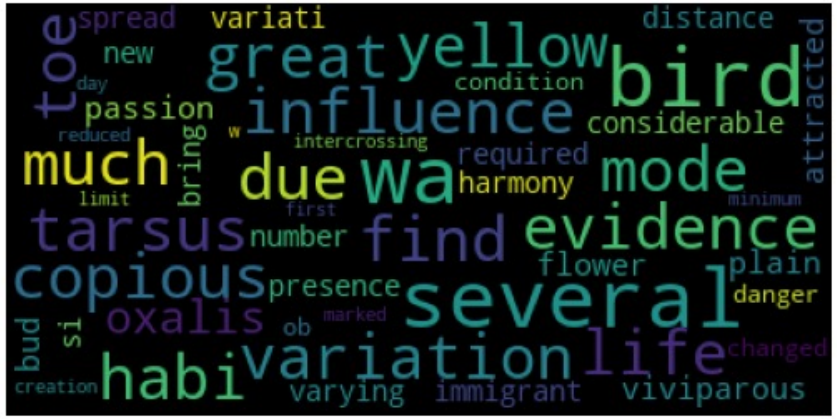






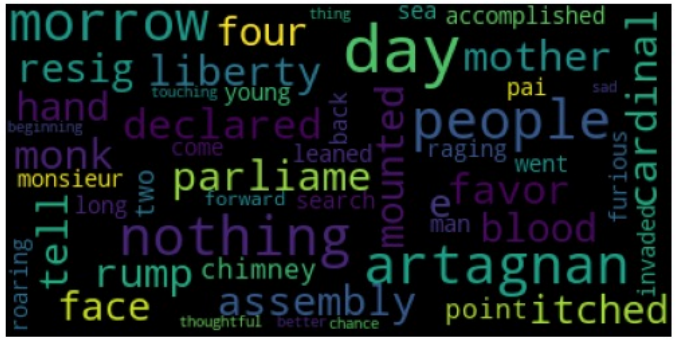
Frequent Words: Wordcloud

50 Frequently used words in the book: EBook of Darwinism



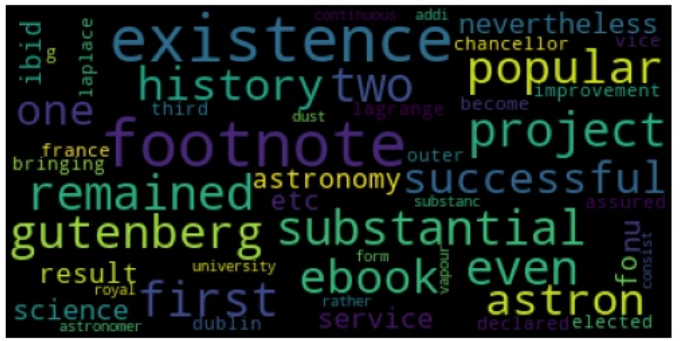
50 Frequently used words in the book: Chaldea

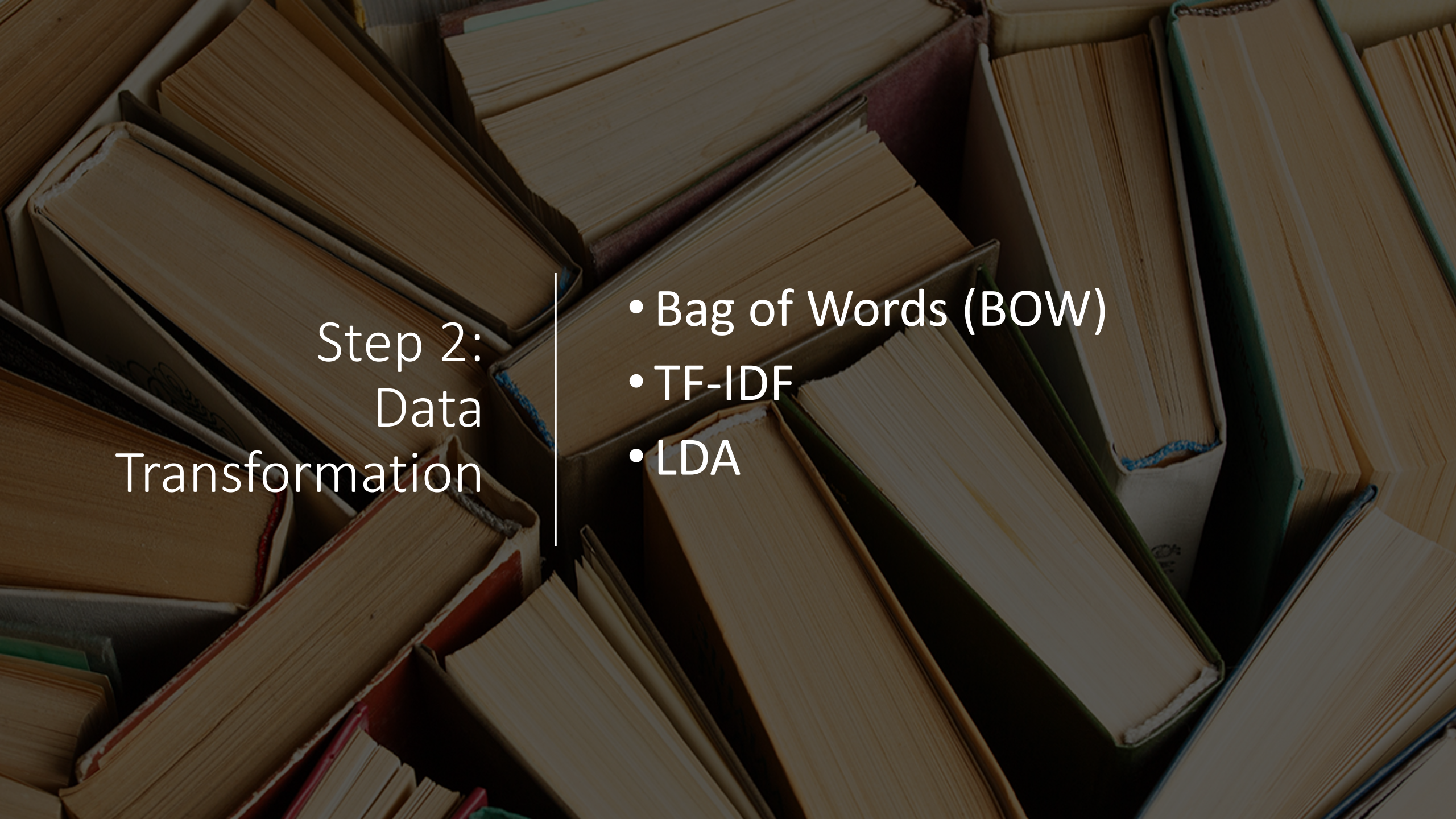
50 Frequently used words in the book: The Vicomte de Bragelonne



50 Frequently used words in the book: A Book About Lawyers

50 Frequently used words in the book: A Popular History of Astronomy During the Nineteenth C
entury





Step 2: Data Transformation

- Bag of Words (BOW)
- TF-IDF
- LDA

BOW:

```
BOW
```

	aaron	abandon	abandoned	abandoning	abandonment	abated	abb	abbe	abbey	abbott	...	zodiacal	zonal	zone	zool	zoologique	zoologist	zoolog
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	0	...	0	0	0	0	0	0	
2	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	0	...	0	0	0	0	0	0	
4	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	
996	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	
998	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	

1000 rows × 16048 columns

TF-IDF:

```
TFIDF_Vector
```

	aaron	abandon	abandoned	abandoning	abandonment	abated	abb	abbe	abbey	abbott	...	zodiacal	zonal	zone	zool	zoologique	zoologist	zoolog
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	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	0.0	0.0	0.0	0.0	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	0.0	0.0	0.0	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	0.0	0.0	0.0	0.0	0.0	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	0.0	0.0	0.0	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	...	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	...	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	...	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	...	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	...	0.0	0.0	0.0	0.0	0.0	0.0	0.

1000 rows × 16048 columns

LDA:

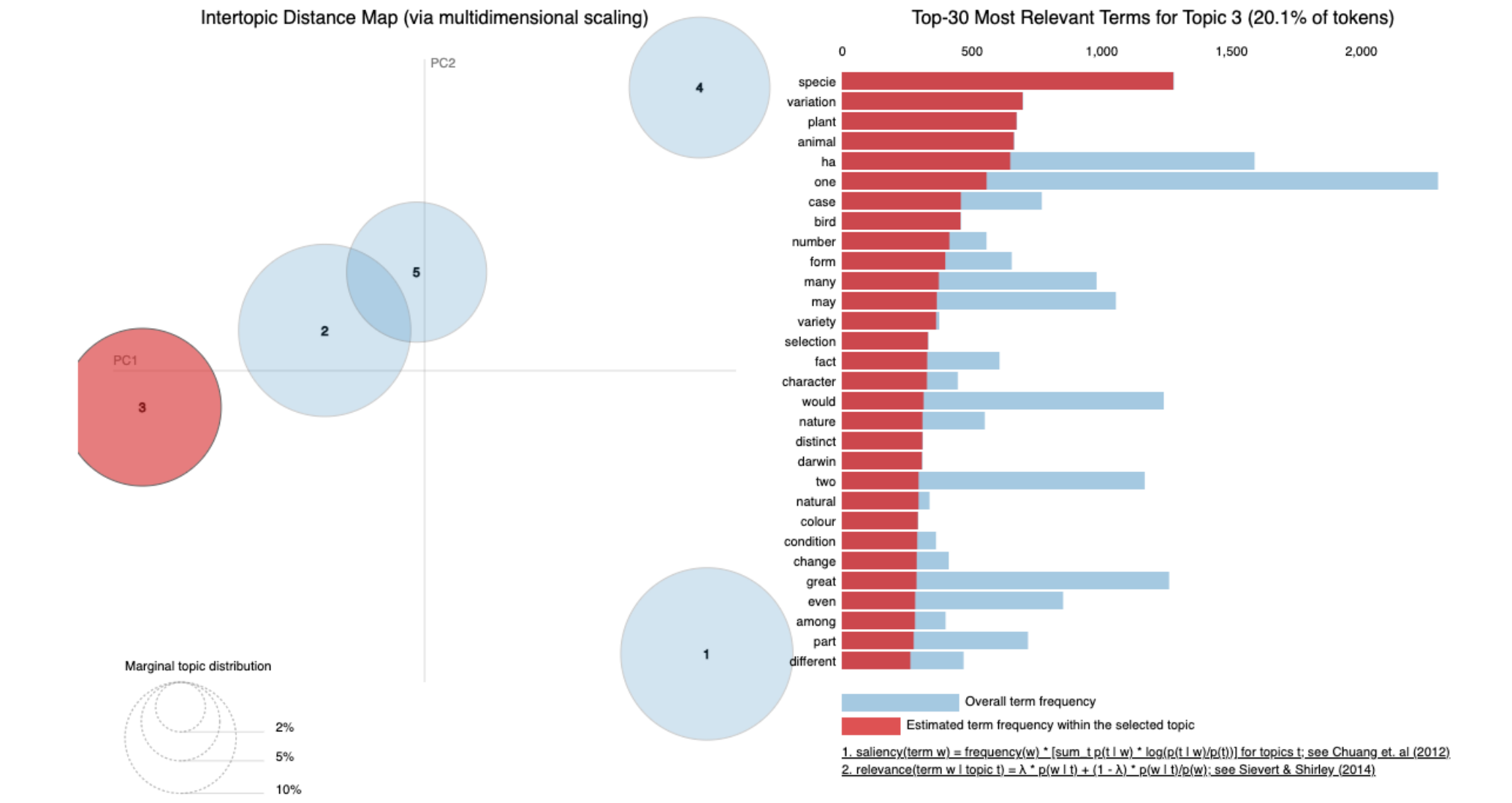
```
LDA.head()
```

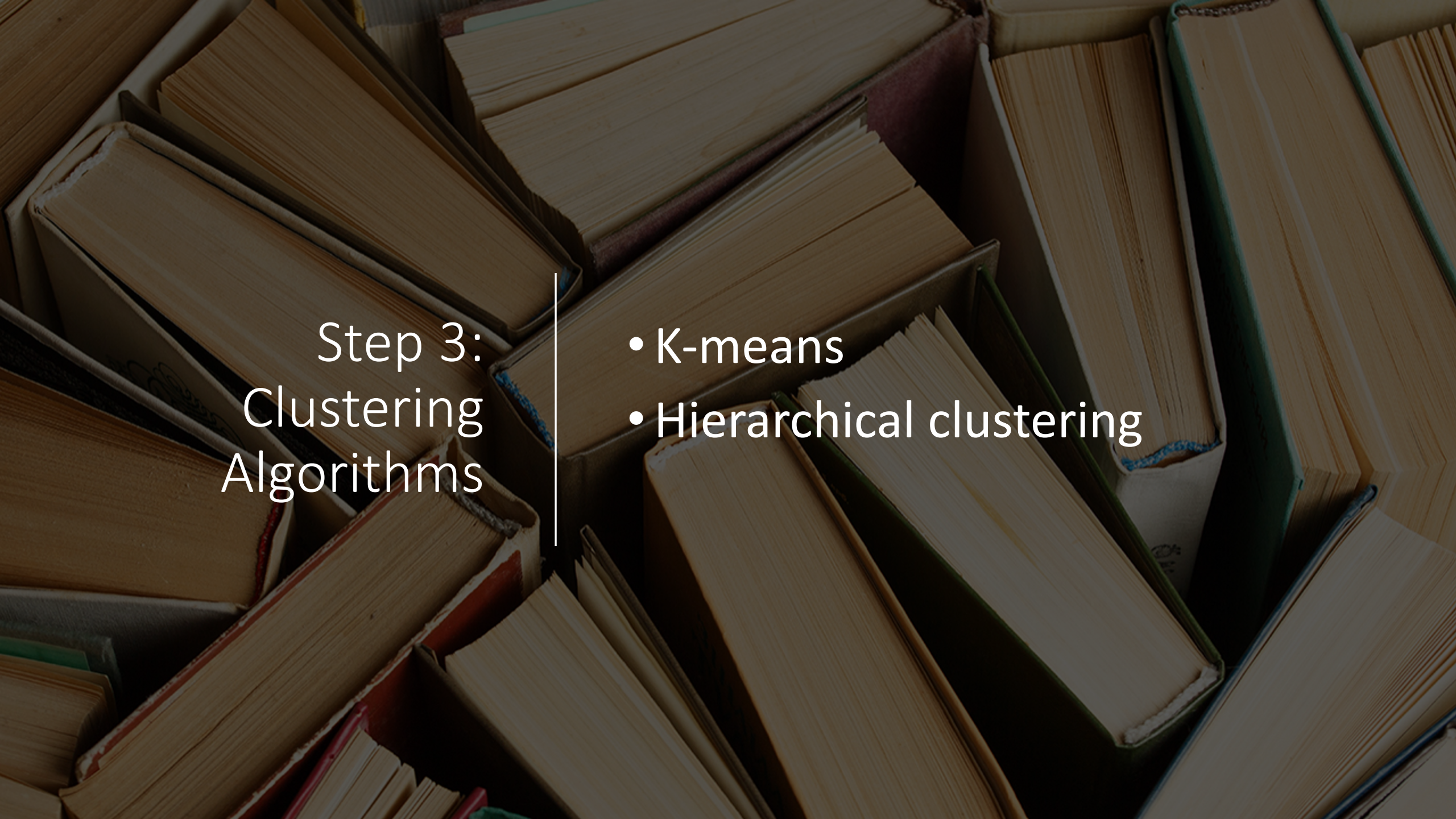
	1	2	3	4	5	res
0	0.121710	6.297770	10.518360	41.755322	91.040962	5
1	128.704269	6.219258	0.131042	15.295912	0.270618	1
2	84.970520	0.204380	6.978431	27.358265	30.191488	1
3	3.745756	0.206495	0.131131	146.006546	0.268842	4
4	81.228470	1.564073	0.131100	51.285492	15.966647	1

```
PredictedWords
```

```
(array([[1.2170985e-01, 6.2977695e+00, 1.0518360e+01, 4.1755322e+01,
         9.1040962e+01],
        [1.2870427e+02, 6.2192578e+00, 1.3104180e-01, 1.5295912e+01,
         2.7061805e-01],
        [8.4970520e+01, 2.0437954e-01, 6.9784307e+00, 2.7358265e+01,
         3.0191488e+01],
        ...,
        [3.1825294e+00, 9.5775023e+00, 1.3119207e-01, 3.2911915e-01,
         1.3765221e+02],
        [1.2156795e-01, 3.0078484e+01, 1.3103145e-01, 1.2030704e+02,
         2.6176342e-01],
        [3.7553685e+00, 1.6282141e+01, 2.9358027e+01, 9.1212921e+01,
         1.0117374e+01]], dtype=float32),
 None)
```

LDA as Topic Modeling:



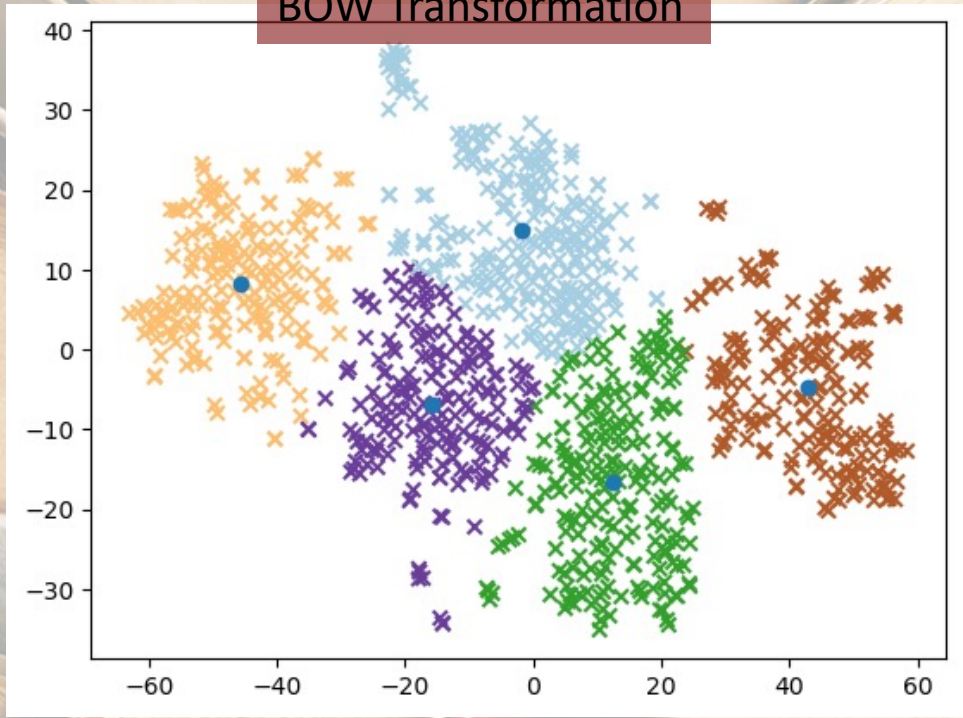


Step 3: Clustering Algorithms

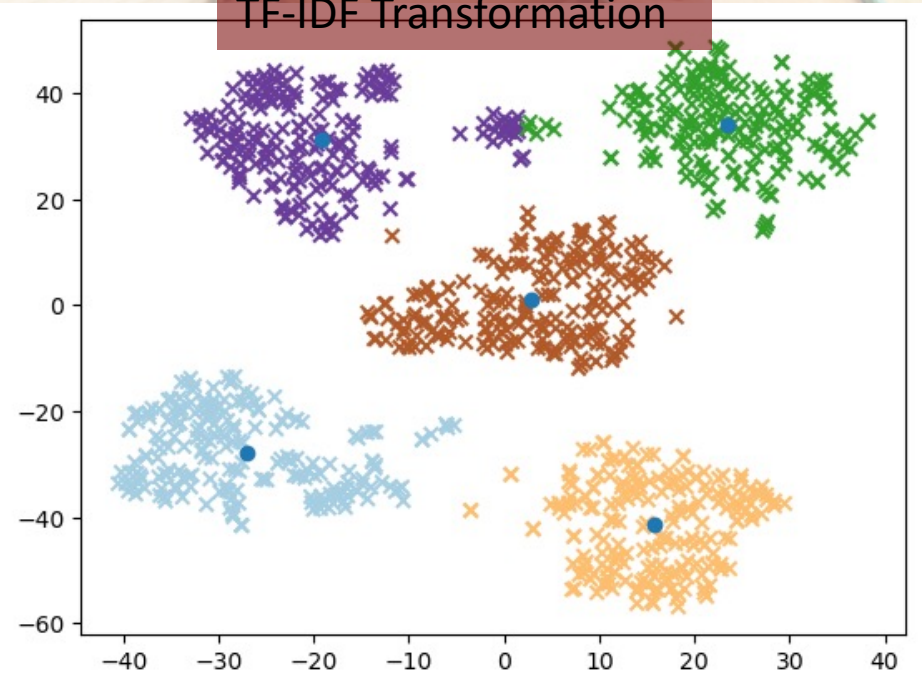
- K-means
- Hierarchical clustering

K-means Clustering:

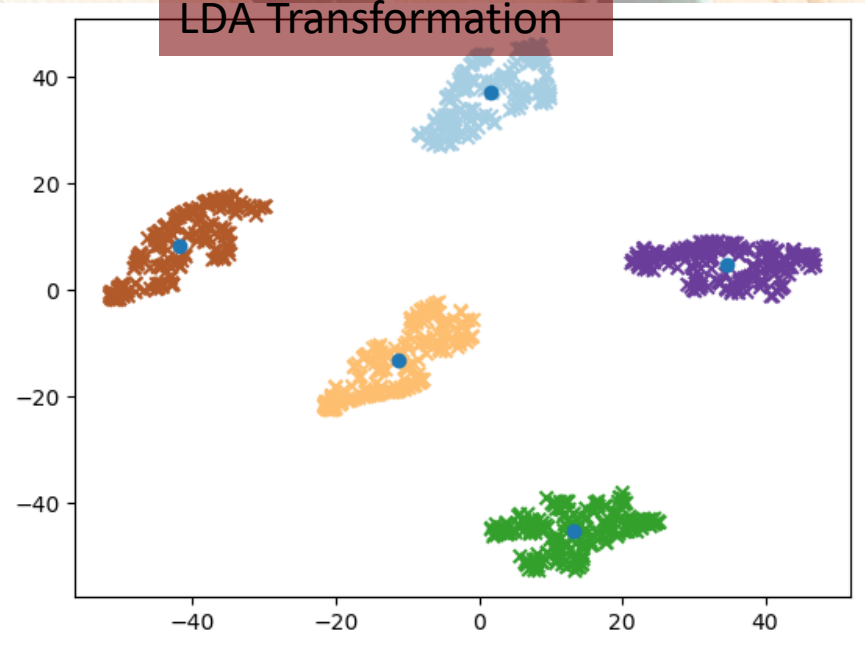
BOW Transformation



TF-IDF Transformation

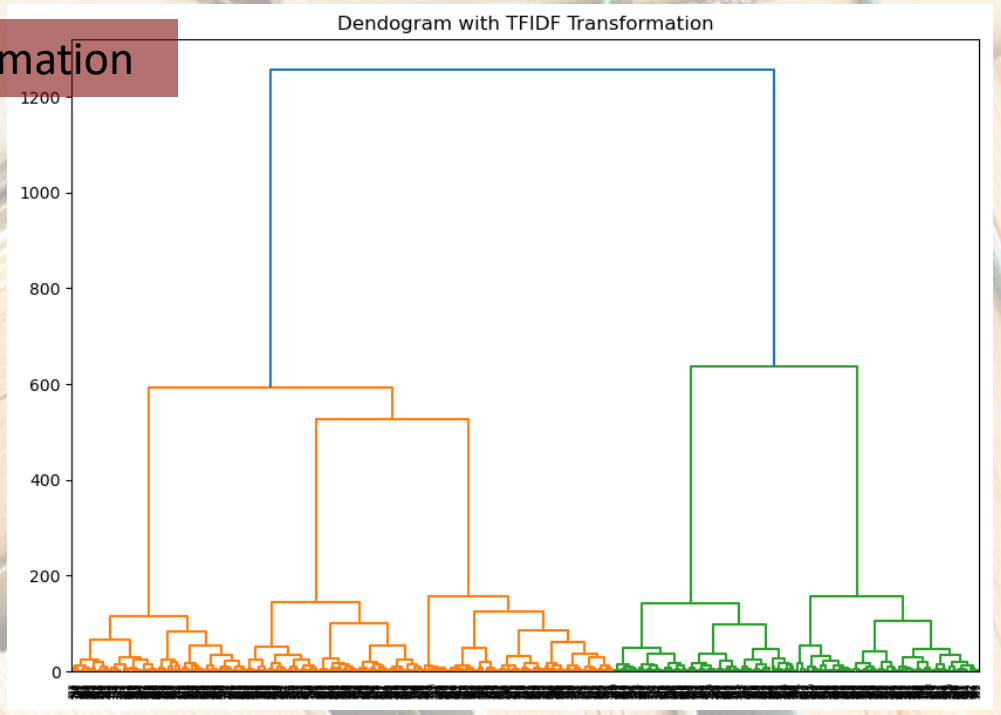


LDA Transformation

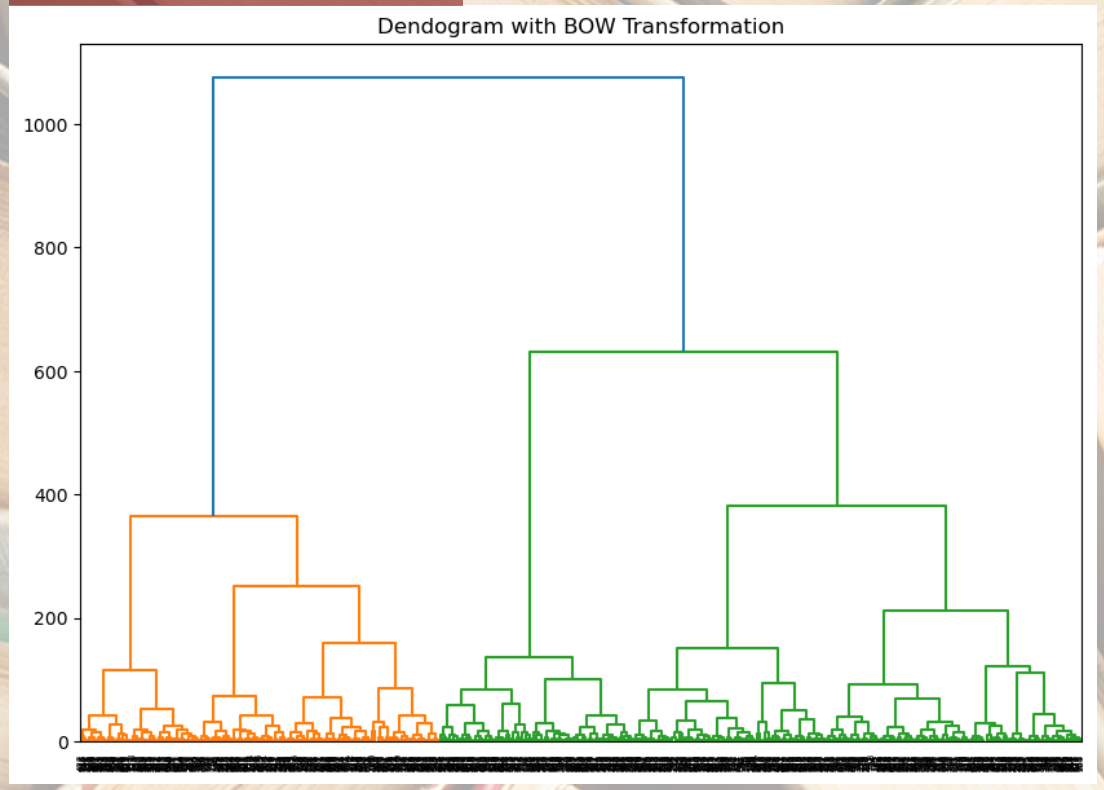


Hierarchical Clustering:

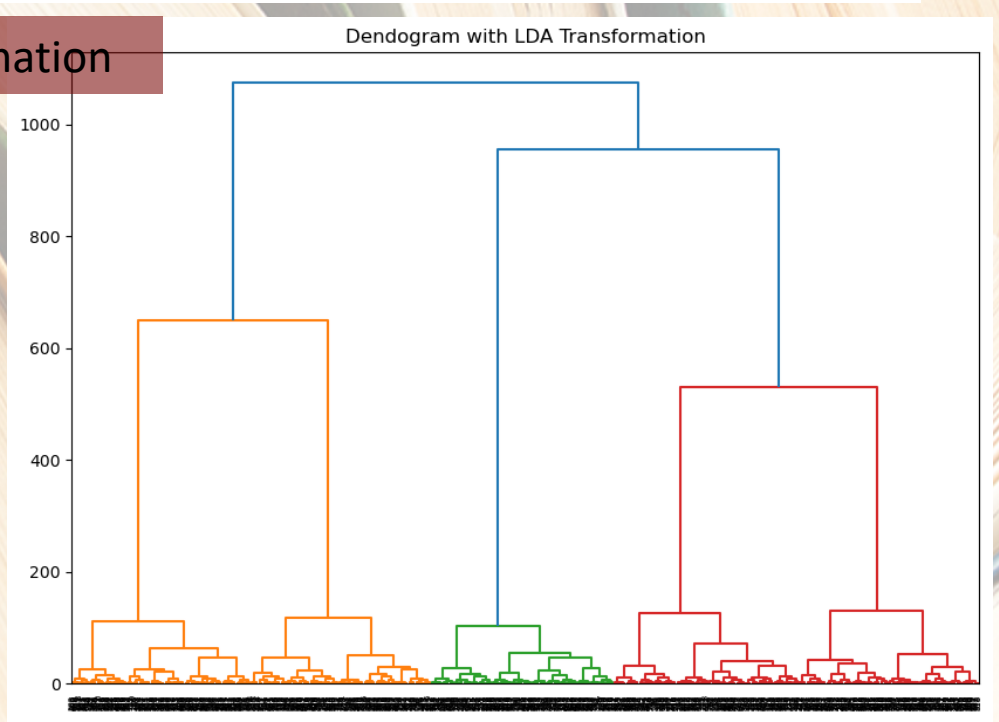
TF-IDF Transformation

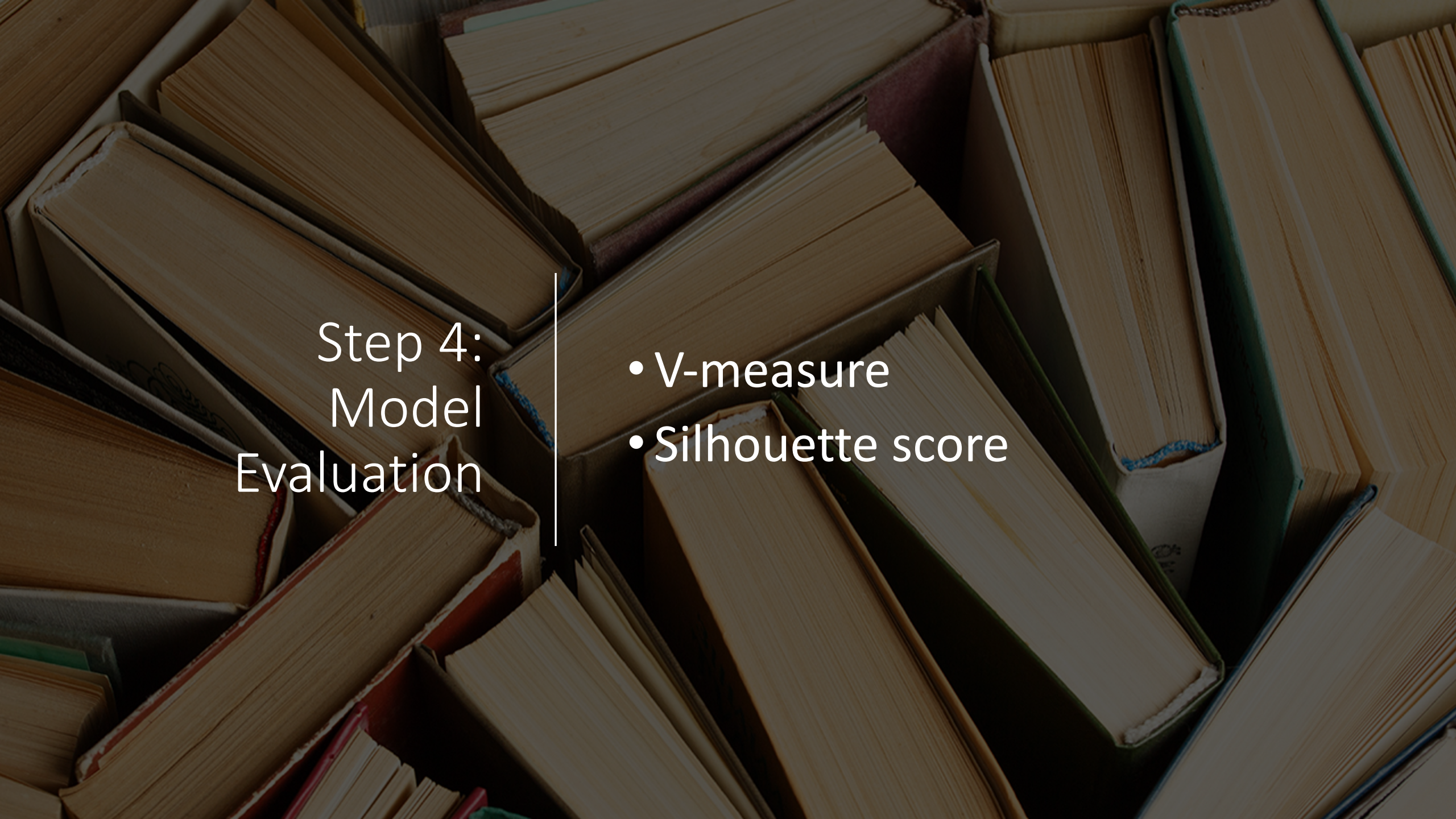


BOW Transformation



LDA Transformation

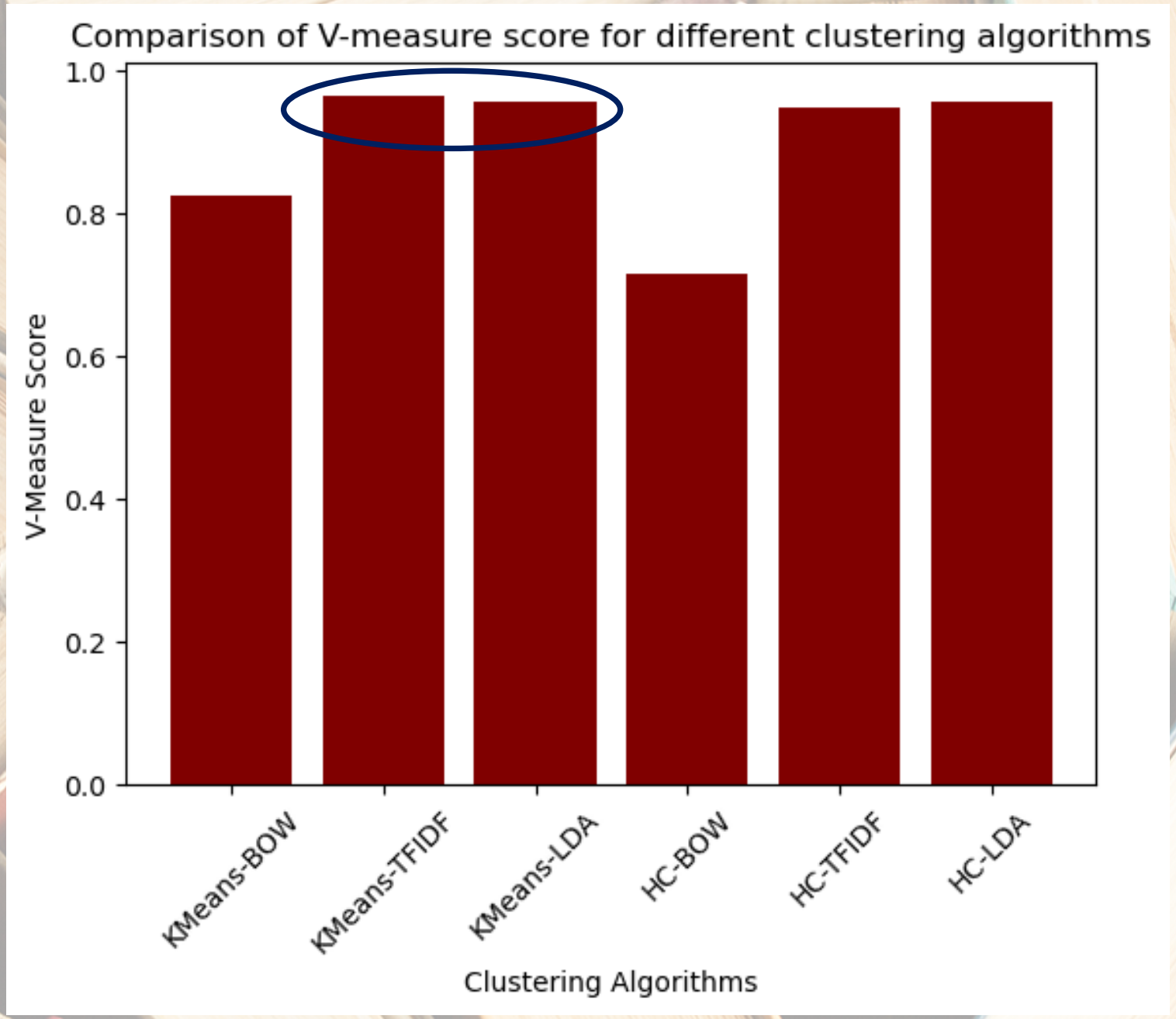




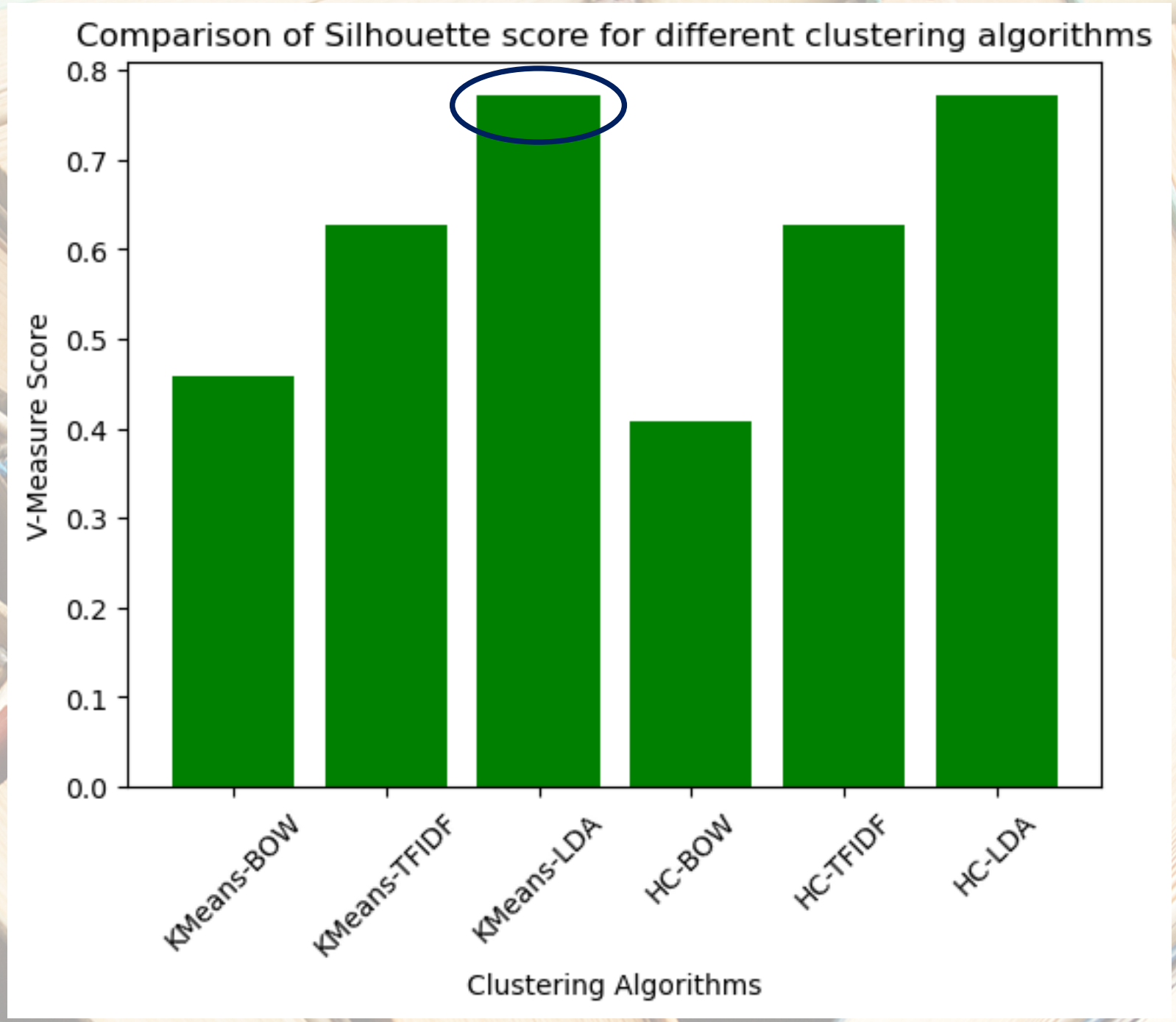
Step 4: Model Evaluation

- V-measure
- Silhouette score

V-measure:

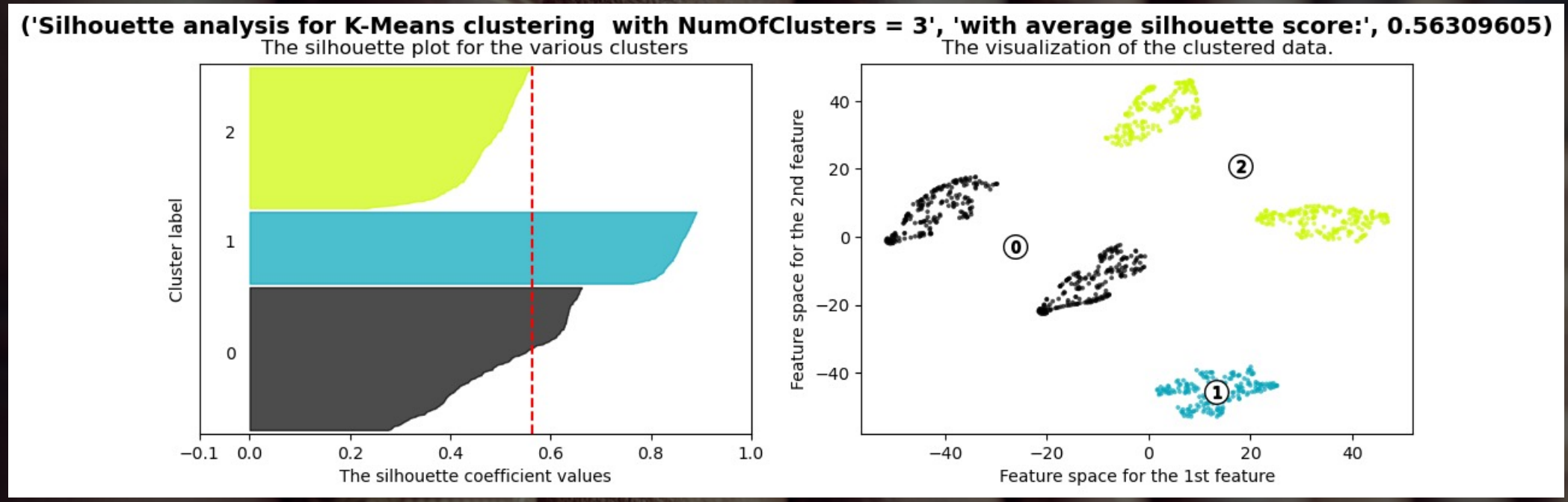

Comparison of V-measure score for different clustering algorithms
V-Measure Score
1.0
0.8
0.6
0.4
0.2
0.0
KMeans-BOW
KMeans-TFIDF
KMeans-LDA
HC-BOW
HC-TFIDF
HC-LDA
Clustering Algorithms

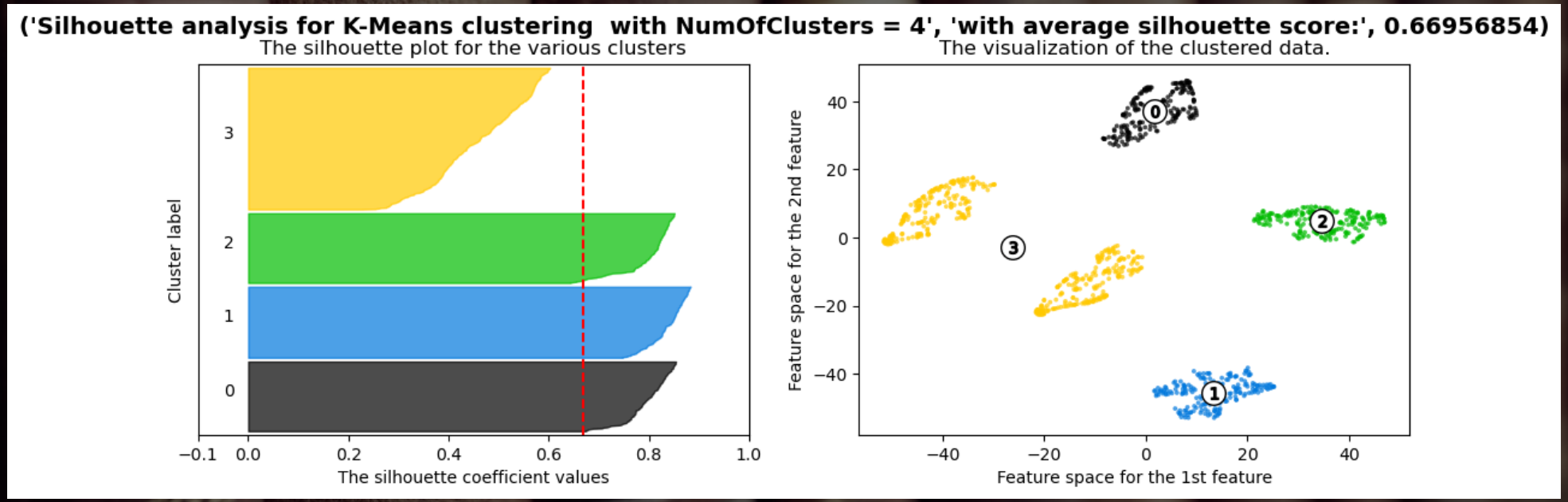
Silhouette score:

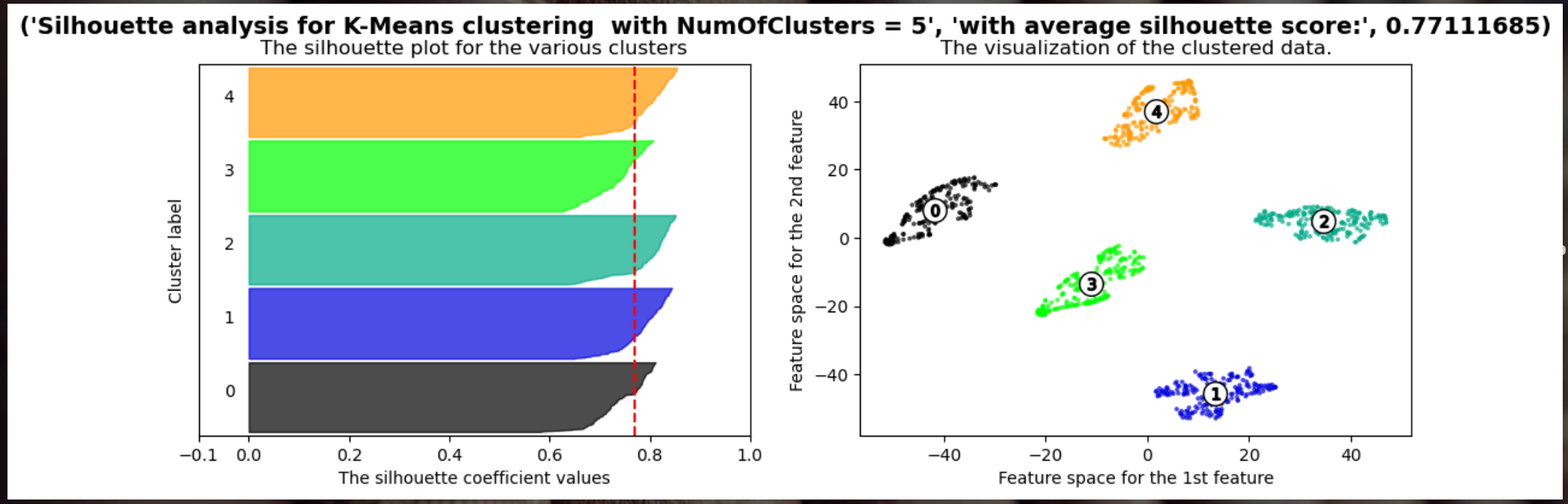


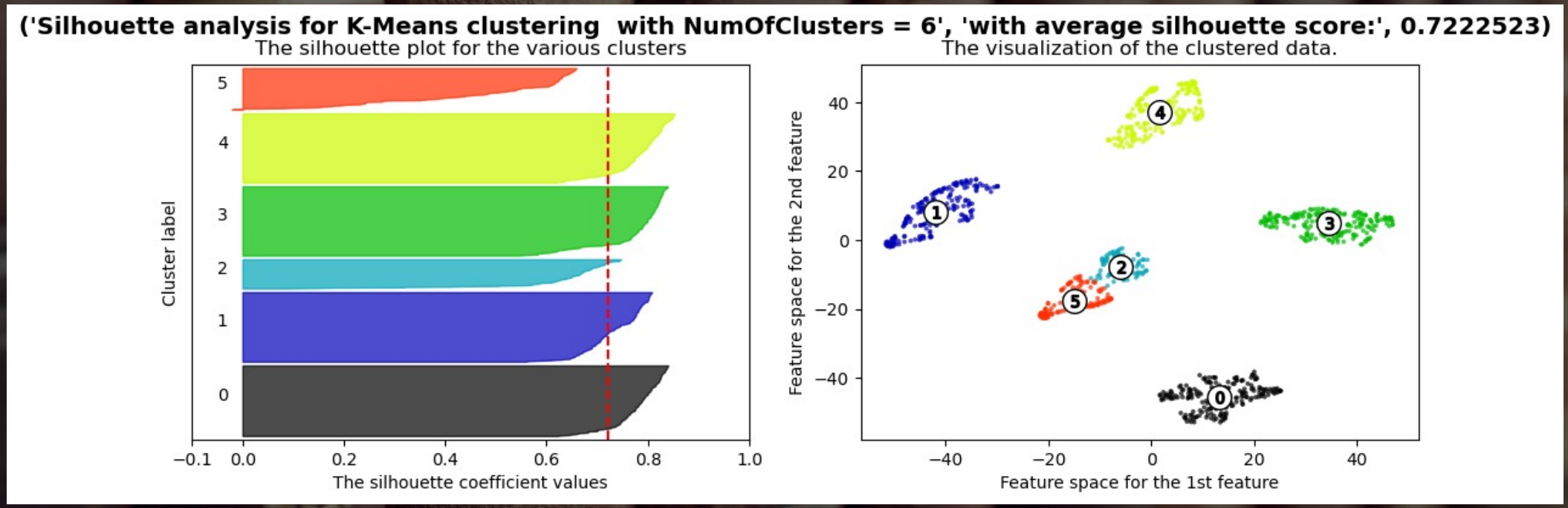

Best Model
K-means on LDA transformed data

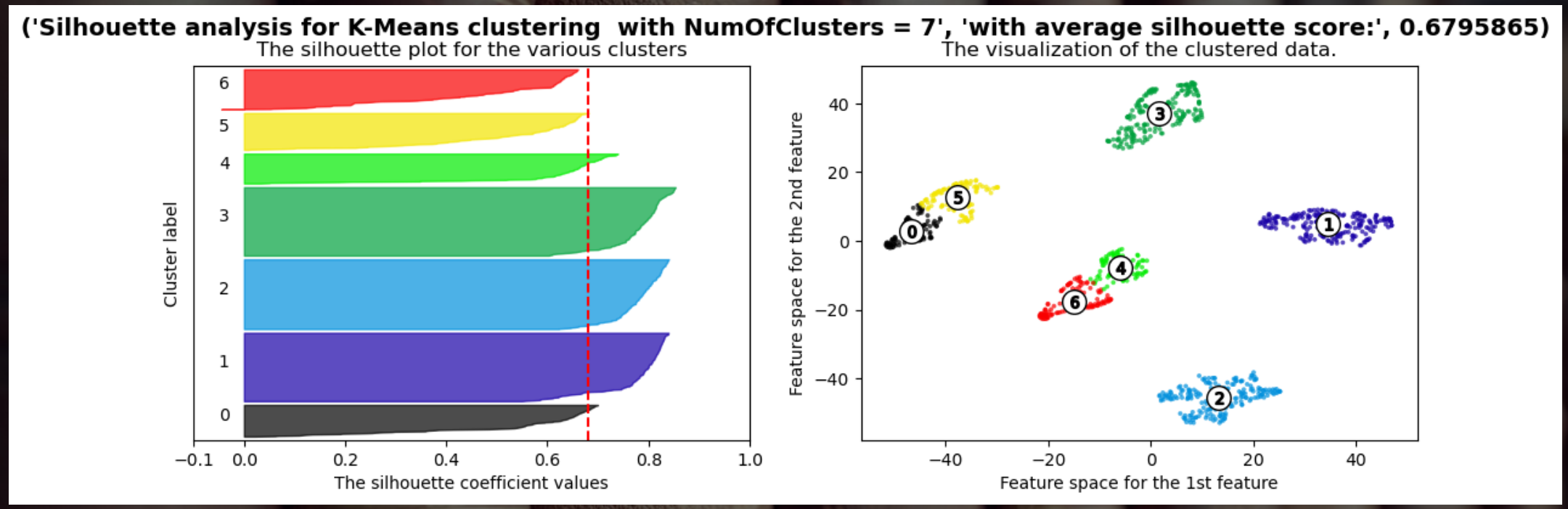

Finding the optimum K
using Silhouette scores

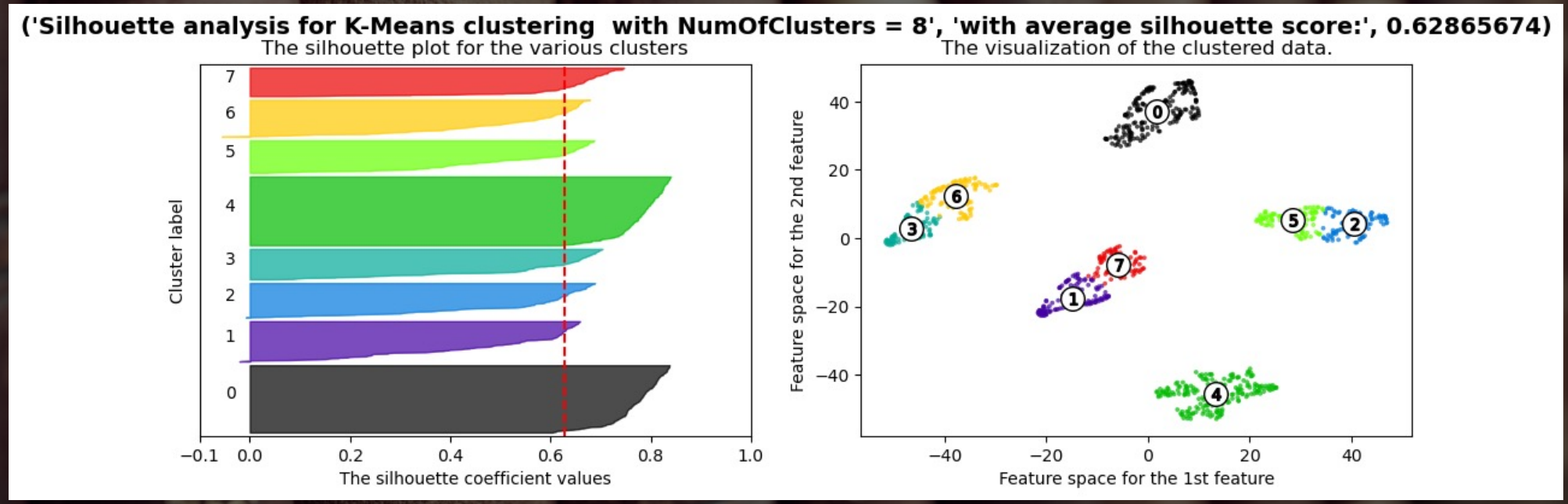

('Silhouette analysis for K-Means clustering with NumOfClusters = 3', 'with average silhouette score:', 0.56309605)
The silhouette plot for the various clusters
Cluster label
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
40
20
0
−20
−40
−40
−20
0
20
40
Feature space for the 1st feature
0
2
1

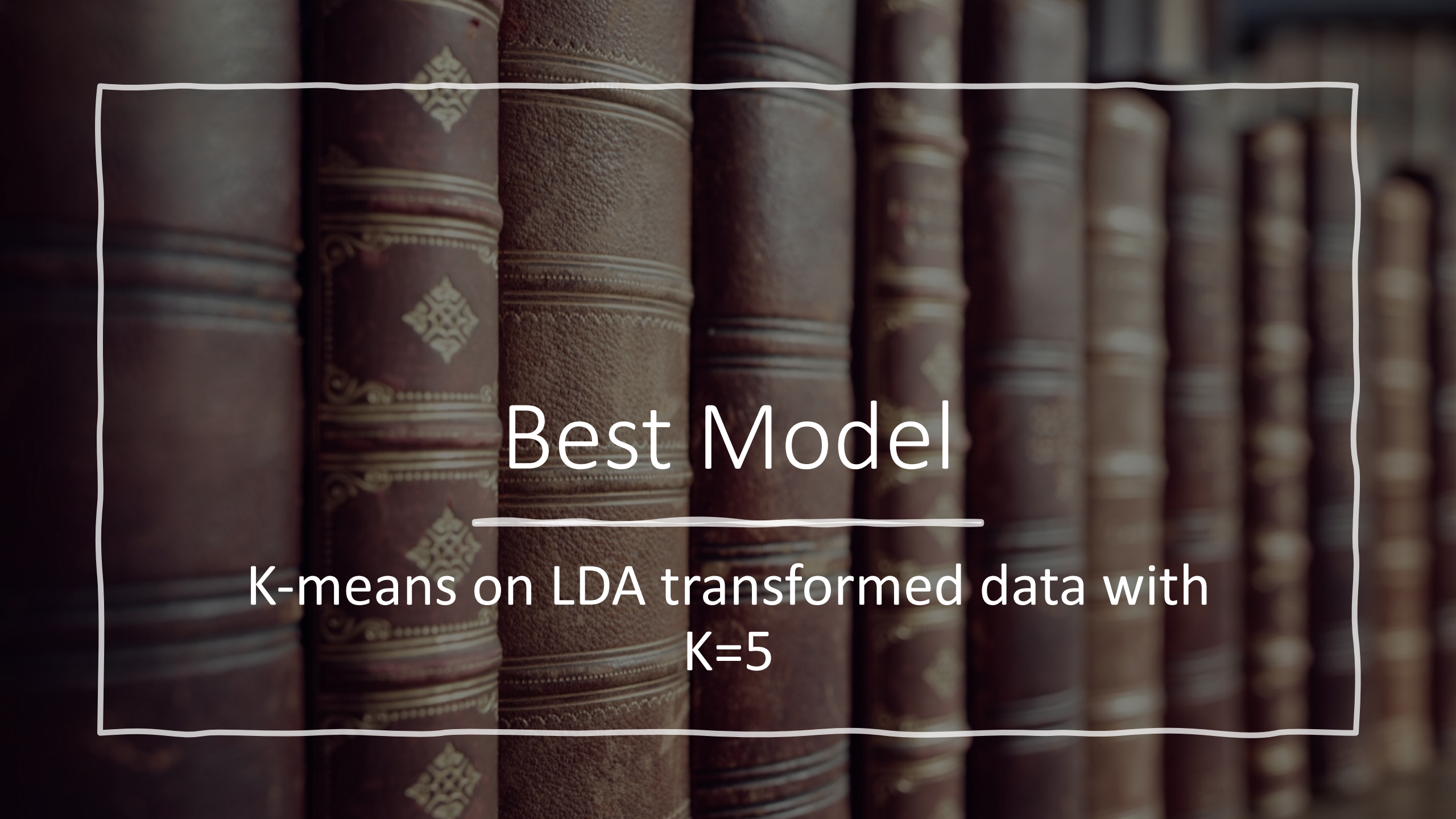

('Silhouette analysis for K-Means clustering with NumOfClusters = 4', 'with average silhouette score:', 0.66956854)
The silhouette plot for the various clusters
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
40
20
0
−20
−40
−40
−20
0
20
40
Feature space for the 1st feature
0
2
3
1


('Silhouette analysis for K-Means clustering with NumOfClusters = 5', 'with average silhouette score:', 0.77111685)
The silhouette plot for the various clusters
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature


('Silhouette analysis for K-Means clustering with NumOfClusters = 6', 'with average silhouette score:', 0.7222523)
The silhouette plot for the various clusters
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature

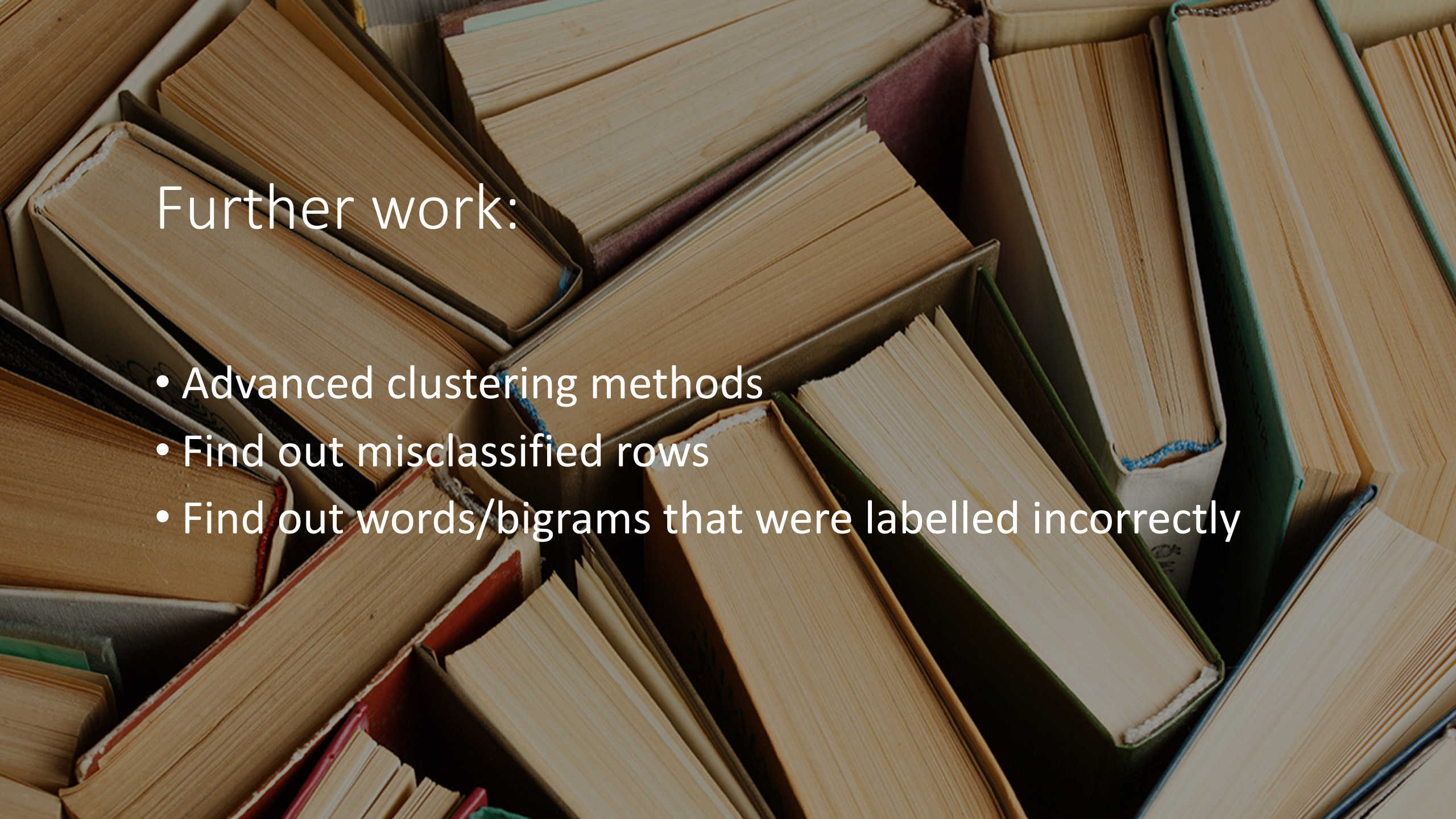

('Silhouette analysis for K-Means clustering with NumOfClusters = 7', 'with average silhouette score:', 0.6795865)
The silhouette plot for the various clusters
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature


('Silhouette analysis for K-Means clustering with NumOfClusters = 8', 'with average silhouette score:', 0.62865674)
The silhouette plot for the various clusters
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature



Best Model

K-means on LDA transformed data with K=5



Further work:

- Advanced clustering methods
- Find out misclassified rows
- Find out words/bigrams that were labelled incorrectly

Thank you