NATURAL LANGUAGE PROCESSING-BASED TEXT CLUSTERING

-Catherine Mariana Philips

TABLE OF CONTENTS

PROBLEM STATEMENT

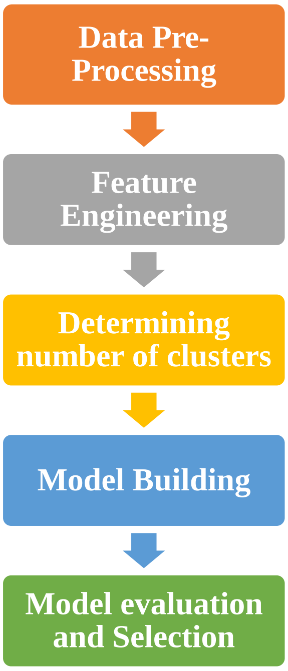
To estimate the number of clusters using a new method of estimation of the number of optimal K's called Depth Difference method and to cluster the given religious texts using different clustering methods.

OBJECTIVE OF THE PROJECT

The dataset given to us consists of sacred religious texts from 8 different books. 4 books are from Asian text and the latter 4 are Biblical text. There are 2 objectives in the project.

1. The traditional techniques such as 'Elbow plot' and 'Silhouette Analysis' do not yield the appropriate number of K values and so we move towards a new method called Depth Difference. We experiment on this method to determine the appropriate number of K's and plot it.
2. Using the number of K's gained from the Depth Difference method, we implement 3 clustering techniques namely **'K-means clustering', 'Spectral clustering', and 'Hierarchical clustering'** to group and cluster the religious text.

PIPELINE FOLLOWED



Data Pre-Processing
Feature Engineering
Determining number of clusters
Model Building
Model evaluation and Selection

ABOUT THE DATA

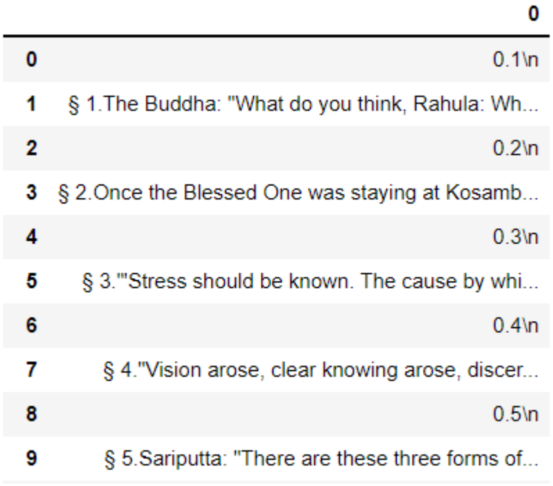
The given dataset consists of religious text of two categories.

1. Asian religious texts
2. Biblical texts

The Asian religious text consists of readings from ‘Buddhism’, ‘Tao Te Ching’, ‘Upanishad’, and ‘Yoga sutra’ and the Biblical text from the books of ‘Wisdom’, ‘Proverbs’, ‘Ecclesiastes’, and ‘Ecclesiasticus’ respectively.

DATA PRE-PROCESSING

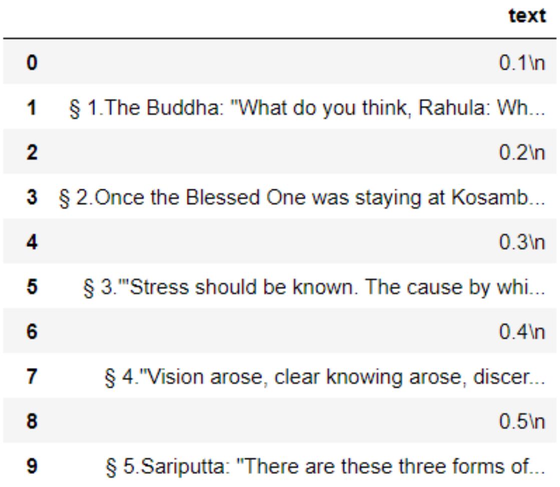
The given data is first loaded as a list. The list is then converted to a data frame.



	0
0	0.1\n
1	§ 1.The Buddha: "What do you think, Rahula: Wh...
2	0.2\n
3	§ 2.Once the Blessed One was staying at Kosamb...
4	0.3\n
5	§ 3."'Stress should be known. The cause by whi...
6	0.4\n
7	§ 4."Vision arose, clear knowing arose, discer...
8	0.5\n
9	§ 5.Sariputta: "There are these three forms of...

1: DataFrame

Since we convert the list to a data frame, the column header is 0 by default. We change it to ‘text’ for easier manipulation.

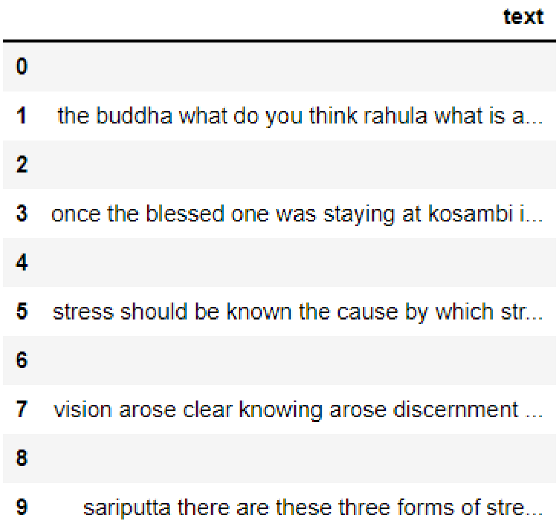


	text
0	0.1\n
1	§ 1.The Buddha: "What do you think, Rahula: Wh...
2	0.2\n
3	§ 2.Once the Blessed One was staying at Kosamb...
4	0.3\n
5	§ 3."'Stress should be known. The cause by whi...
6	0.4\n
7	§ 4."Vision arose, clear knowing arose, discer...
8	0.5\n
9	§ 5.Sariputta: "There are these three forms of...

2:After changing column header

a) CLEANING THE DATA

- Cleaning the data consists of
  - ✓ removing the special character ‘§’
  - ✓ removing punctuations such as '!"#$%&\'() *+, -. /:;<=>?@[\\]^_`{|}~'
  - ✓ removing numerical data
  - ✓ removing newline
  - ✓ converting the text to lowercase



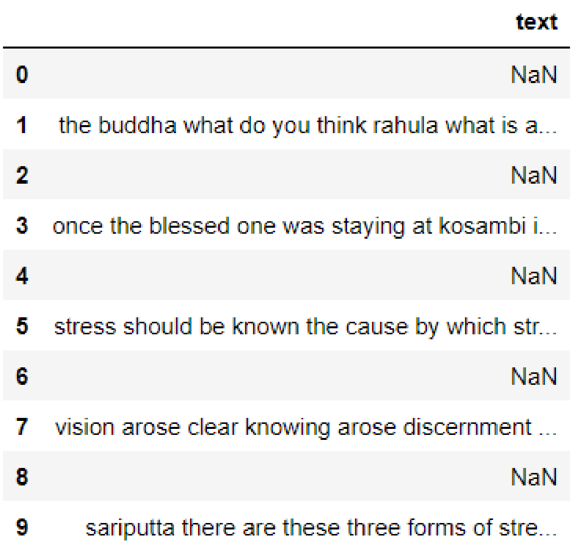
	text
0	
1	the buddha what do you think rahula what is a...
2	
3	once the blessed one was staying at kosambi i...
4	
5	stress should be known the cause by which str...
6	
7	vision arose clear knowing arose discernment ...
8	
9	sariputta there are these three forms of stre...

3:After cleaning the data

- After the data is cleaned, few blank rows can be seen. To drop the blank rows, we will first fill the blank spaces NaN values.

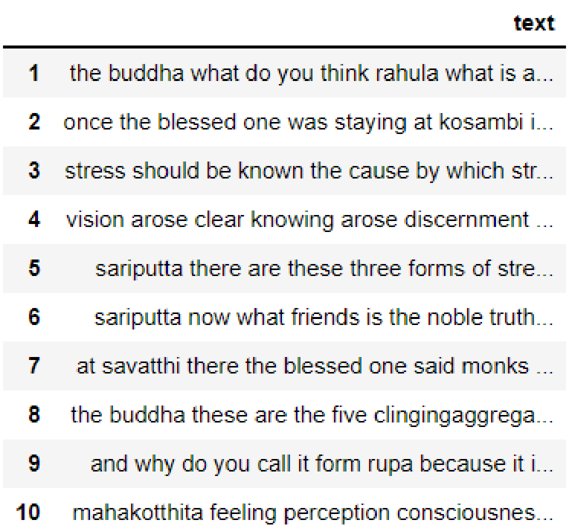
- The NaN values are dropped. The index values are now 1,3,5, etc.

- The indices are changed to get a continuous series of rows.



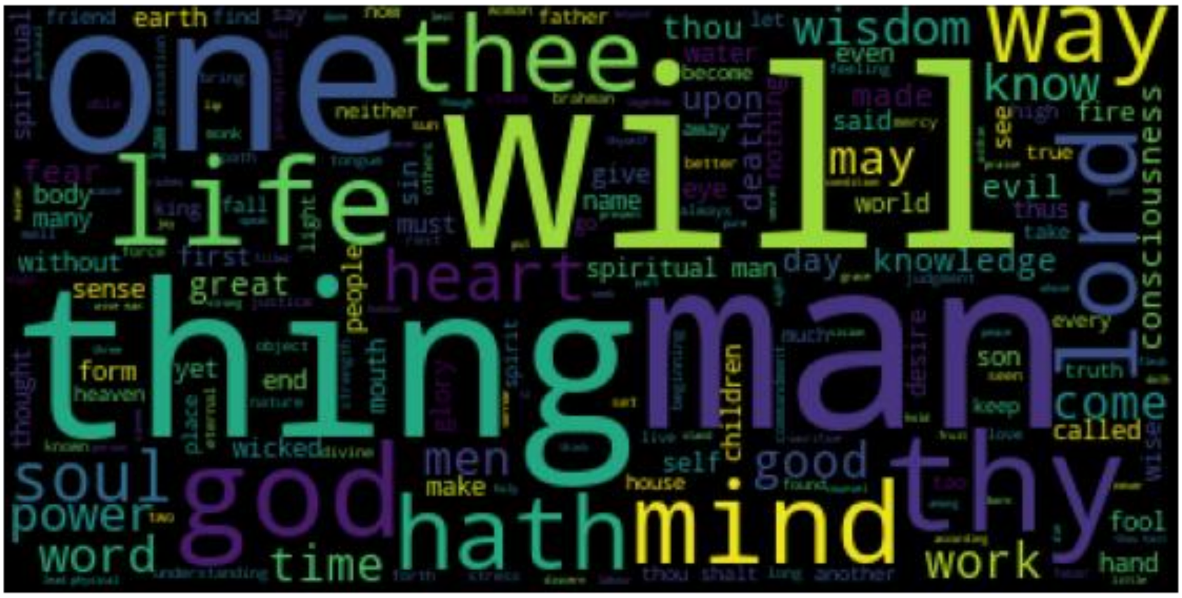
	text
0	NaN
1	the buddha what do you think rahula what is a...
2	NaN
3	once the blessed one was staying at kosambi i...
4	NaN
5	stress should be known the cause by which str...
6	NaN
7	vision arose clear knowing arose discernment ...
8	NaN
9	sariputta there are these three forms of stre...

4:After filling with NaN



	text
1	the buddha what do you think rahula what is a...
2	once the blessed one was staying at kosambi i...
3	stress should be known the cause by which str...
4	vision arose clear knowing arose discernment ...
5	sariputta there are these three forms of stre...
6	sariputta now what friends is the noble truth...
7	at savatthi there the blessed one said monks ...
8	the buddha these are the five clingingaggrega...
9	and why do you call it form rupa because it i...
10	mahakotthita feeling perception consciousnes...

5:DataFrame after removing NaN rows



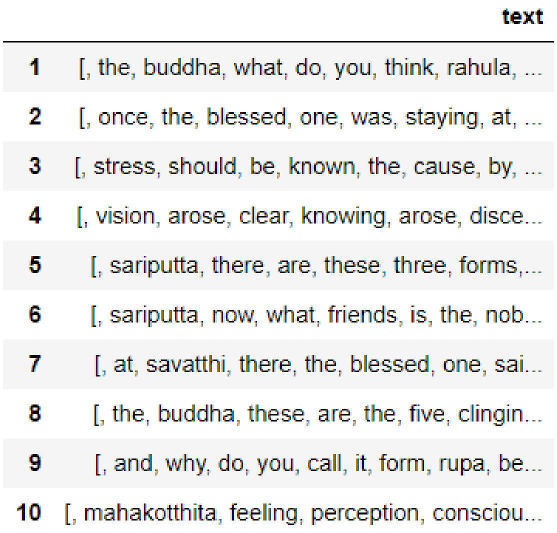
6:WordCloud

- From the word cloud, we can see that the words **one, thee, way, will, life, thing, man, lord, god, hath,** and **thy** appear a greater number of times in the given data.

b) <u>TOKENIZING THE DATA</u>

- Tokenization is the process of breaking down a phrase, sentence, paragraph, or even an entire text document into smaller components like individual words or phrases.
- These smaller units are called tokens.

- The given data is tokenized into tokens.



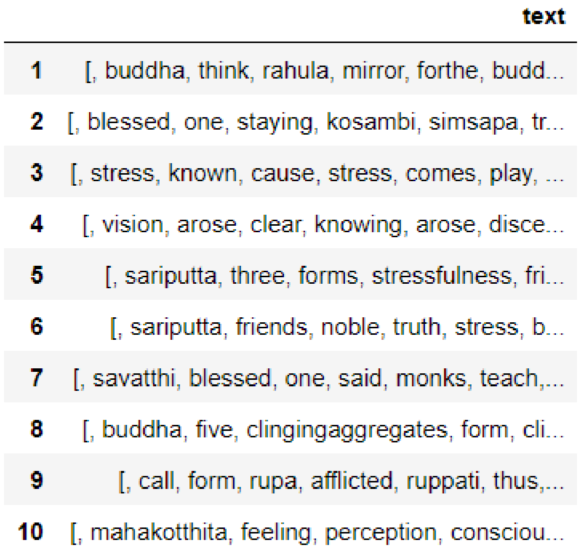
	text
1	[, the, buddha, what, do, you, think, rahula, ...
2	[, once, the, blessed, one, was, staying, at, ...
3	[, stress, should, be, known, the, cause, by, ...
4	[, vision, arose, clear, knowing, arose, disce...
5	[, sariputta, there, are, these, three, forms,...
6	[, sariputta, now, what, friends, is, the, nob...
7	[, at, savatthi, there, the, blessed, one, sai...
8	[, the, buddha, these, are, the, five, clingin...
9	[, and, why, do, you, call, it, form, rupa, be...
10	[, mahakotthita, feeling, perception, consciou...

7:Tokenized Data

c) <u>REMOVING STOP WORDS</u>

- Stop words are words that are typically filtered out before a text is processed. These are the most common words in any language (articles, prepositions, pronouns, conjunctions, and so on), and they don't add anything to the text.

- Examples of a few stop words in English are “the”, “a”, “an”, “so”, “what”.



	text
1	[, buddha, think, rahula, mirror, forthe, budd...
2	[, blessed, one, staying, kosambi, simsapa, tr...
3	[, stress, known, cause, stress, comes, play, ...
4	[, vision, arose, clear, knowing, arose, disce...
5	[, sariputta, three, forms, stressfulness, fri...
6	[, sariputta, friends, noble, truth, stress, b...
7	[, savatthi, blessed, one, said, monks, teach,...
8	[, buddha, five, clingingaggregates, form, cli...
9	[, call, form, rupa, afflicted, ruppati, thus,...
10	[, mahakotthita, feeling, perception, consciou...

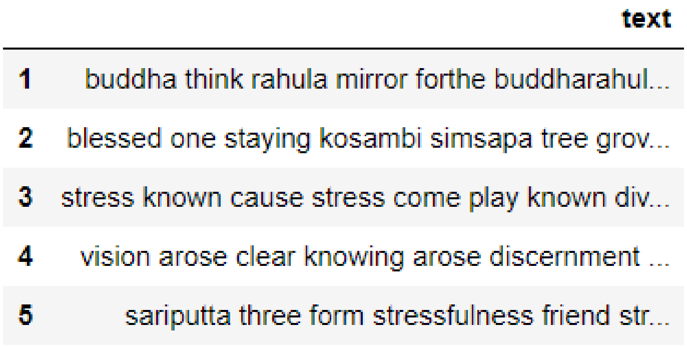
8:After removing stop words

- We remove the low-level information from our text by deleting these terms, allowing us to focus more on the crucial information. In other words, we may say that removing such phrases has no negative impact on the model we are training for our task.

- Since there are fewer elements involved in the training, eliminating stop words reduces the dataset size and thus reduces training time.

d) LEMMATIZATION

- Lemmatization is the process of combining a word's several inflected forms into a single item that may be studied.
- Lemmatization is similar to stemming, but it gives the words context.
- Because lemmatization performs morphological examination of the words, it is preferable over stemming.
- The given tokens are lemmatized to its root word using the WordNet Lemmatizer and are joined to as sentences.

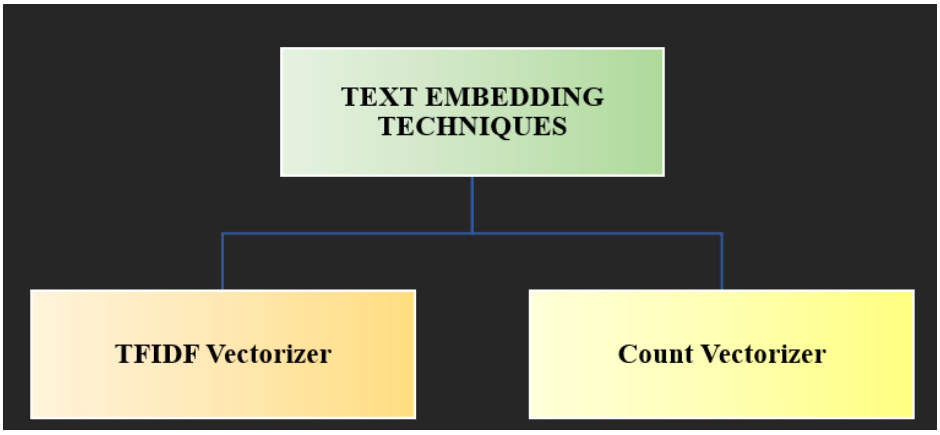


	text
1	buddha think rahula mirror forthe buddharahul...
2	blessed one staying kosambi simsapa tree grov...
3	stress known cause stress come play known div...
4	vision arose clear knowing arose discernment ...
5	sariputta three form stressfulness friend str...

9:After lemmatization

e) EMBEDDING TECHNIQUES

- Text embedding techniques are used to convert and represent words mathematically.



TEXT EMBEDDING TECHNIQUES

TFIDF Vectorizer

Count Vectorizer

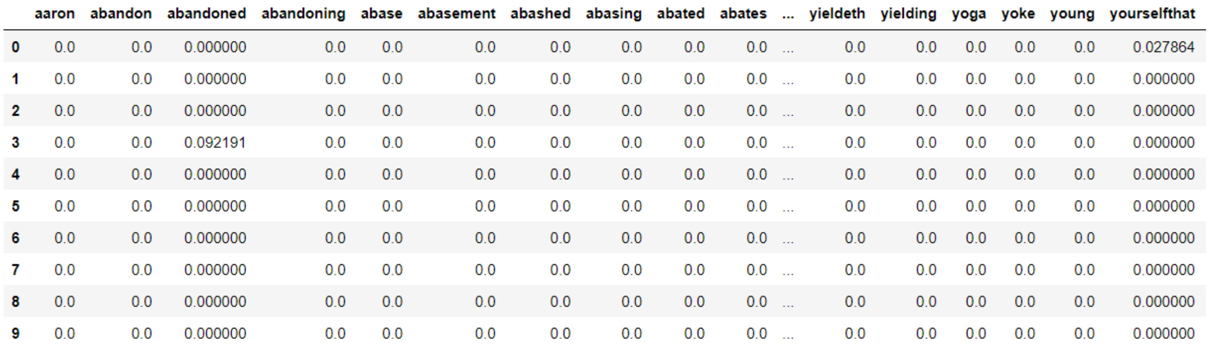
10:Embedding techniques used in the project

i) TFIDF Vectorizer

- TF-IDF is a statistical method for determining the mathematical importance of words in documents.

- The TfidfVectorizer() is used to perform tfidf vectorization.

- The tfidf data now has 589 rows and 7581 columns.



	aaron	abandon	abandoned	abandoning	abase	abasement	abashed	abasing	abated	abates	...	yieldeth	yielding	yoga	yoke	young	yourselfthat
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	0.0	0.0	0.0	0.027864
1	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000
2	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000
3	0.0	0.0	0.092191	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000
4	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000
5	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000
6	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000
7	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000
8	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000
9	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000

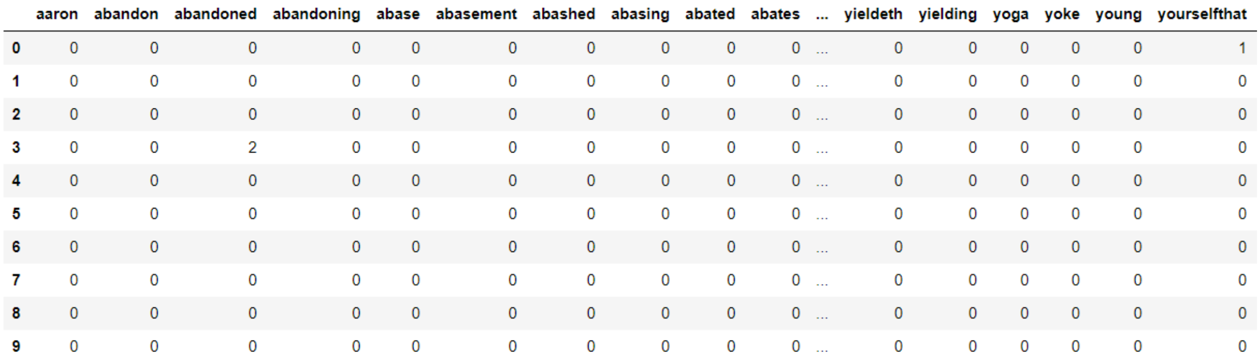
11:TFIDF Vectorized data

ii) Count Vectorizer

- Count Vectorizer converts a given text into a vector based on the frequency (count) of each word that appears in the text.

- The CountVectorizer() is used to perform count vectorization.

- The count vectorizer data now has 589 rows and 7581 columns.



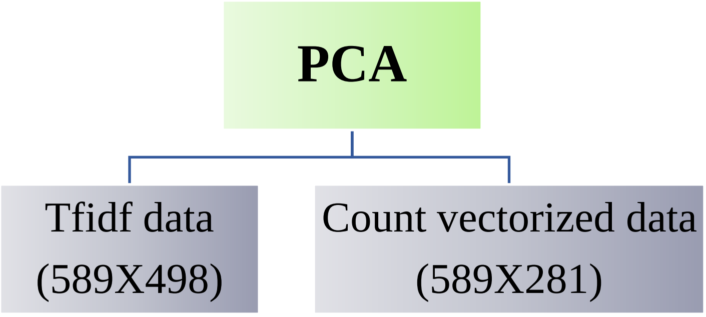
	aaron	abandon	abandoned	abandoning	abase	abasement	abashed	abasing	abated	abates	...	yieldeth	yielding	yoga	yoke	young	yourselfthat
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
3	0	0	2	0	0	0	0	0	0	0	...	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

12:Count Vectorized data

TF-IDF is better than Count Vectorizer because it not only considers the frequency of words in the text, but also their relative significance.

FEATURE ENGINEERING

- The technique of choosing, altering, and transforming raw data into features that may be employed in machine learning is known as feature engineering.
- The feature engineering technique used in this project is **Principal Component Analysis.**
- PCA, or principal component analysis, is a statistical process that allows us to summarize information from enormous amounts of data. It is a dimensionality reduction technique that finds the significant trends in data.
- We perform PCA on both tfidf vectorized and count vectorized data.

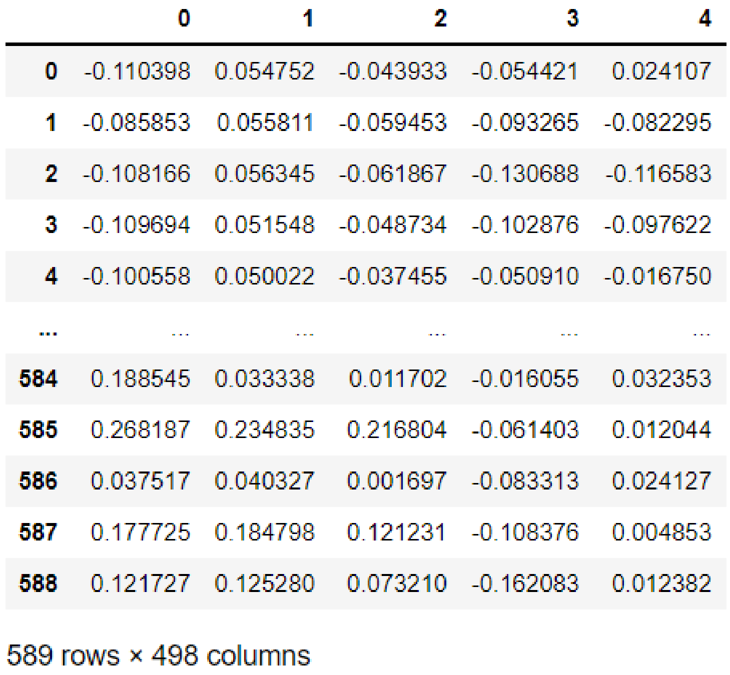


PCA

Tfidf data (589X498)

Count vectorized data (589X281)

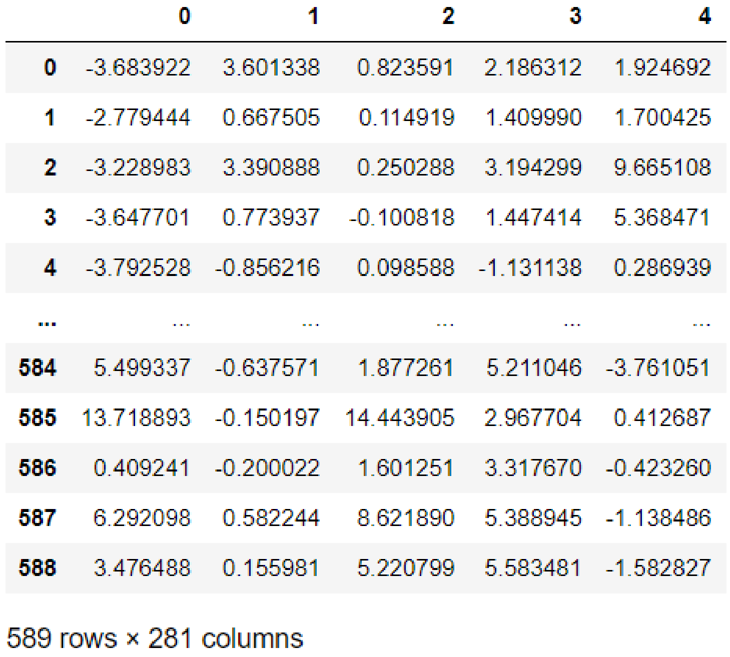
- After applying PCA on both tfidf vectorized and count vectorized data, we get a reduced data.
- The original data had 589 rows and 7581 columns and they are reduced to 589 rows and 498 columns (Tfidf data) and 589 rows and 281 columns(count vectorized data)



	0	1	2	3	4
0	-0.110398	0.054752	-0.043933	-0.054421	0.024107
1	-0.085853	0.055811	-0.059453	-0.093265	-0.082295
2	-0.108166	0.056345	-0.061867	-0.130688	-0.116583
3	-0.109694	0.051548	-0.048734	-0.102876	-0.097622
4	-0.100558	0.050022	-0.037455	-0.050910	-0.016750
...	...	...	...	...	...
584	0.188545	0.033338	0.011702	-0.016055	0.032353
585	0.268187	0.234835	0.216804	-0.061403	0.012044
586	0.037517	0.040327	0.001697	-0.083313	0.024127
587	0.177725	0.184798	0.121231	-0.108376	0.004853
588	0.121727	0.125280	0.073210	-0.162083	0.012382

589 rows × 498 columns

TFIDF data after PCA



	0	1	2	3	4
0	-3.683922	3.601338	0.823591	2.186312	1.924692
1	-2.779444	0.667505	0.114919	1.409990	1.700425
2	-3.228983	3.390888	0.250288	3.194299	9.665108
3	-3.647701	0.773937	-0.100818	1.447414	5.368471
4	-3.792528	-0.856216	0.098588	-1.131138	0.286939
...	...	...	...	...	...
584	5.499337	-0.637571	1.877261	5.211046	-3.761051
585	13.718893	-0.150197	14.443905	2.967704	0.412687
586	0.409241	-0.200022	1.601251	3.317670	-0.423260
587	6.292098	0.582244	8.621890	5.388945	-1.138486
588	3.476488	0.155981	5.220799	5.583481	-1.582827

589 rows × 281 columns

Count Vectorizer data after PCA

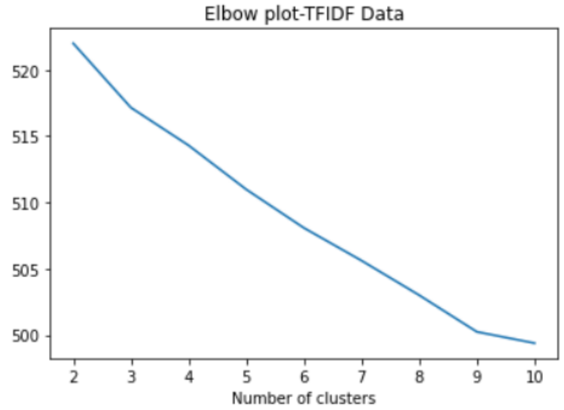
DETERMINING THE NUMBER OF CLUSTERS

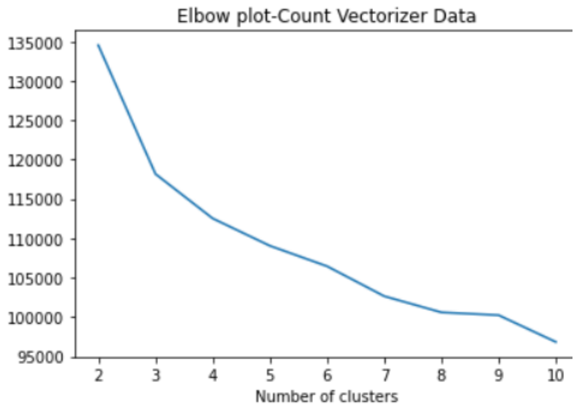
The number of K's needed for clustering are determined using 3 methods in this project.

1. Elbow plot
2. Silhouette Analysis
3. Depth Difference method

1) ELBOW PLOT

- The elbow method is a methodology for figuring out how many clusters there are in a data collection.
- The method consists of plotting the explained variance against the number of clusters and selecting the elbow of the curve as the number of clusters.

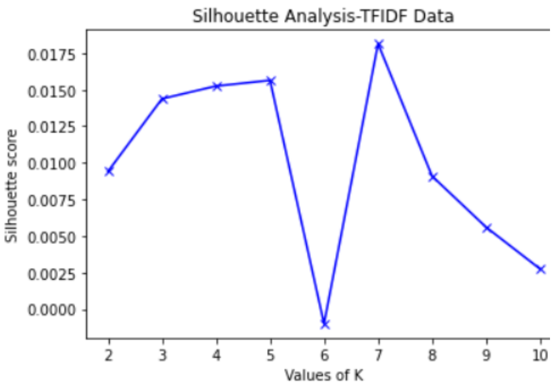


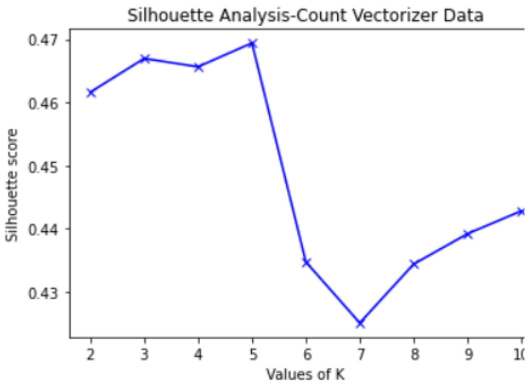


- From the elbow plots, we can see that they do not give the correct number of clusters i.e., 8
- We proceed to the next method called Silhouette Analysis to determine the number of clusters.

2) SILHOUETTE ANALYSIS

- The separation distance between the generated clusters can be studied using silhouette analysis.
- The silhouette plot shows how close each point in one cluster is to points in neighboring clusters, allowing you to visually examine metrics such as the number of clusters.
- The major difference between elbow and silhouette scores is that elbow only calculates the Euclidean distance whereas silhouette takes into account variables such as variance, skewness, high-low differences, etc.





- The silhouette analysis plots show the optimal number of clusters as 7 and 5 respectively for the tfidf and count vectorizer data.
- These plots do not show the optimal number of clusters and so we move on to a new method called as “Depth Difference”

3) DEPTH DIFFERENCE

- The optimal number of clusters (k) in a dataset is estimated using a new method termed depth difference (DeD) which uses data depth.
- Before performing actual clustering, the DeD technique calculates the k parameter.
- To finalize the ideal value of k, we define depth inside clusters, depth across clusters, and depth difference.
- **Data Depth:** The median in a multidimensional dataset, which is the maximum depth in the dataset
- The deepest point in the dataset will be the point with the greatest depth, which will be determined by the **Mahalanobis depth function.**
- The following are calculated.
  - ✓ **DM**: The deepest point in the dataset is the depth median.

    DM = max(Di)
  - ✓ **DW**: Depth within cluster

    The depth median of each cluster is represented as DM^k . Hence,

    DM^k = max(D^k $_i$)

    The mean difference between the depth within a cluster and the depth median is denoted by △k, is as follows

$$\triangle^k = \frac{1}{n_k} \sum_{i \in C_k} |(D_i^k - DM^k)|$$

The depth within cluster (DW) is defined as the average of △k of k clusters and is calculated as shown below

$$\mathrm{DW} = \frac{1}{k} \sum_{i=1}^{k} (\triangle^i)$$

- ✓ DB: Depth between clusters

The mean difference between the depths within the dataset and the depth median is given as follows

$$\triangle = \frac{1}{n}\sum_{i=1}^{n} |(D_i - DM)|$$

The depth between cluster is defined as the difference between △ and DW, and it is defined as follows

$$DB = \triangle - DW$$

- **Depth Difference (DeD)**

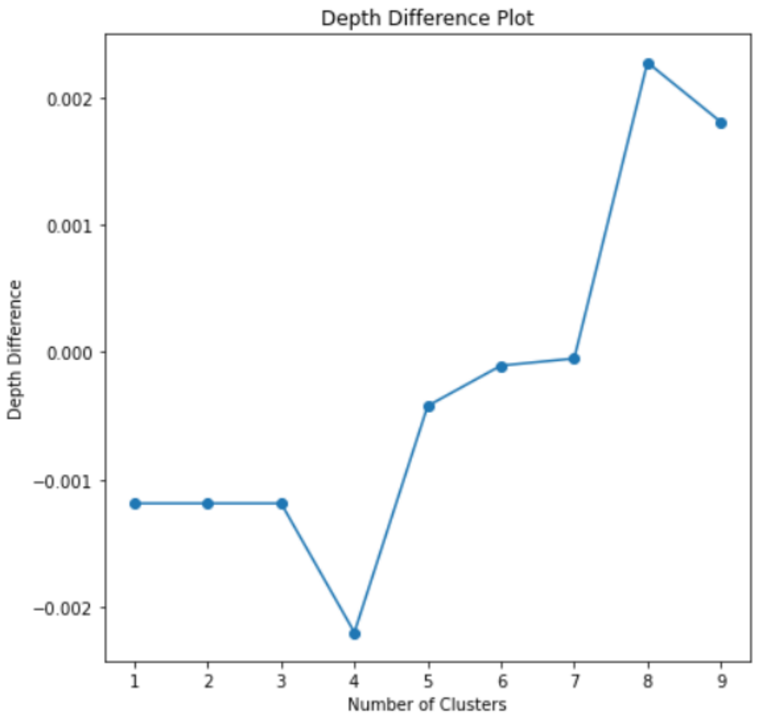
  The depth difference (DeD) is calculated by finding the difference between depth within cluster (DW) and depth between cluster (DB).

$$DeD = DW - DB$$

- **Optimal k**

  The optimal k is the maximum value of DeD.

After finding different values of DeD, the values are plotted and the peak value is taken as the optimal number of K's.



From the plot, we obtain the optimal number of K's i.e 8. Using this, we build the clustering models.

NEED FOR DED METHOD

- The depth difference method is unaffected by the use of unrelated variables.
- It is also unaffected by the presence of high-variance clusters.

- Traditional techniques such as elbow plot and silhouette score select k values by running a clustering algorithm over it with a set of various k values determined by the user.
- DeD, on the contrary, iterates over a dataset with a range of k values to arrive at the optimal number of k clusters, and DeD effectively treats the prevalence of high variance clusters.

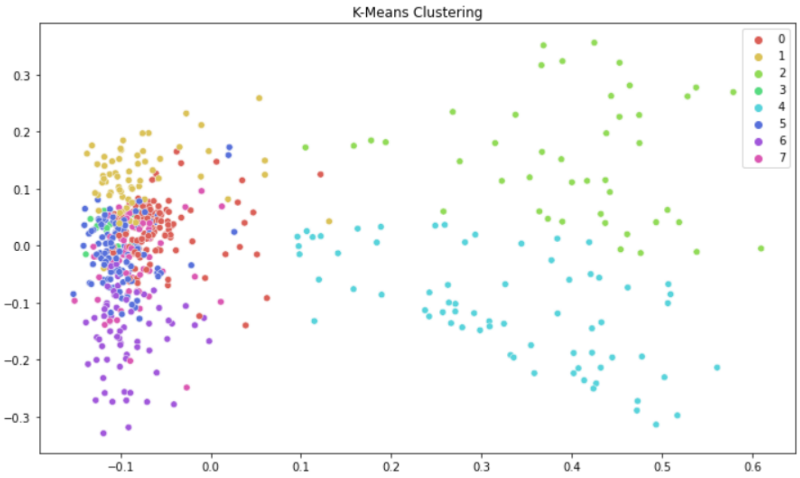
MODEL BUILDING

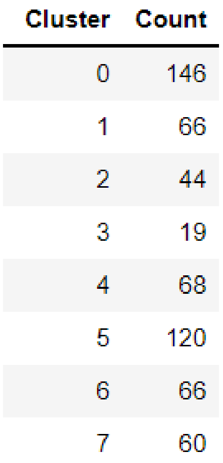
- Clustering is a form of unsupervised learning technique.
- Unsupervised learning is a technique for extracting references from datasets that contain input data but no labelled answers.
- Clustering is the process of partitioning a population or set of data points into several groups so that data in the same group are more similar to each other and dissimilar to data in other groups.
- The clustering techniques used in this project are,
  - ✓ K-means clustering
  - ✓ Spectral clustering
  - ✓ Hierarchical clustering

1) K-MEANS CLUSTERING

- It's the simplest unsupervised learning algorithm for clustering problems.
- The K-means algorithm divides a set of n observations into k groups, with each observation assigned to the cluster with the closest mean.
- K-means clustering algorithm works in three steps.

  - ✓ Select the k values.
  - ✓ Initialize the centroids.
  - ✓ Select the group and find the average.



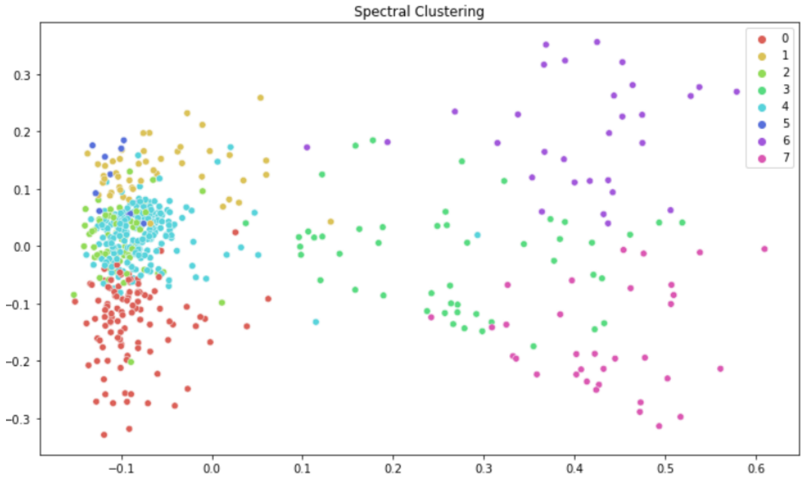


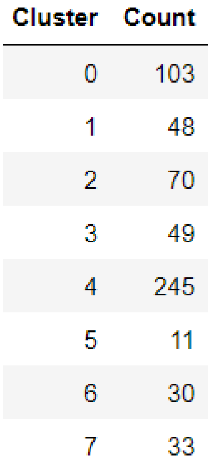
Cluster	Count
0	146
1	66
2	44
3	19
4	68
5	120
6	66
7	60

- The given text data are clustered into 8 clusters and plotted.
- It can be noted that cluster 0 has 146 data points and cluster 3 has 19 data points.

2) SPECTRAL CLUSTERING

- Spectral clustering is a graph partitioning problem.
- Spectral clustering is used to group data that is linked but not always packed or clustered inside convex bounds.
- There are three steps to spectral clustering:
  - ✓ Create a graph of similarity
  - ✓ Transform the data into a two-dimensional space.
  - ✓ Form clusters



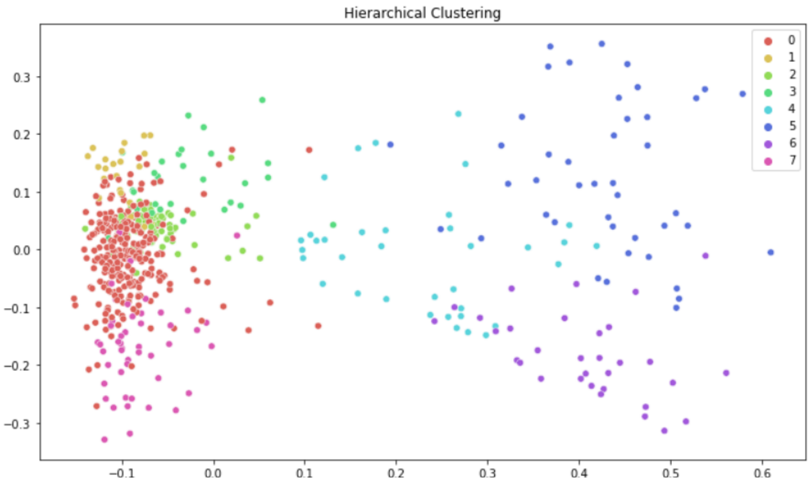


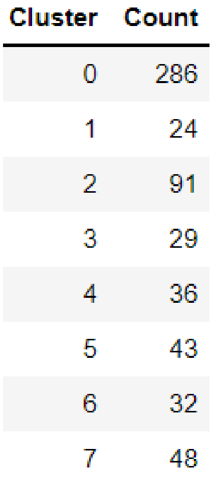
Cluster	Count
0	103
1	48
2	70
3	49
4	245
5	11
6	30
7	33

- The given text data are clustered into 8 clusters and plotted.
- It can be noted that cluster 4 has 245 data points and cluster 5 has 11 data points.

3) HIERARCHICAL CLUSTERING

- Another unsupervised learning approach, hierarchical clustering, is used to group together unlabeled data points with comparable features.
- In hierarchical clustering, each observation is viewed as a separate cluster. The following two steps are then repeated.
  - ✓ Finds the two clusters that are the closest together
  - ✓ The two most comparable clusters are combined. This iterative process is performed until all of the clusters are combined.



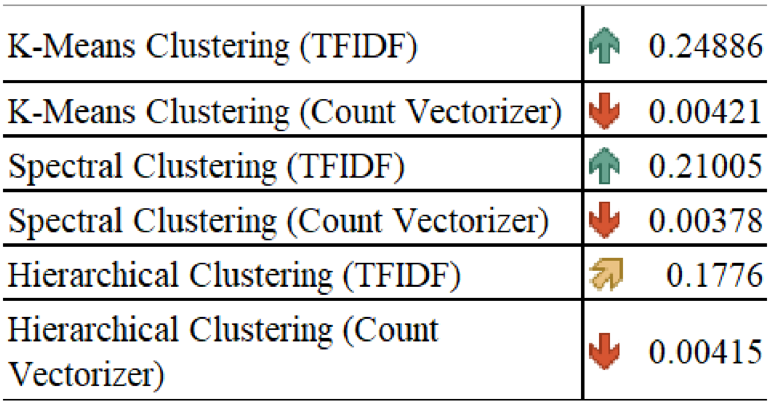


Cluster	Count
0	286
1	24
2	91
3	29
4	36
5	43
6	32
7	48

- The given text data are clustered into 8 clusters and plotted.
- It can be noted that cluster 0 has 286 data points and cluster 1 has 24 data points.

MODEL EVALUATION AND SELECTION

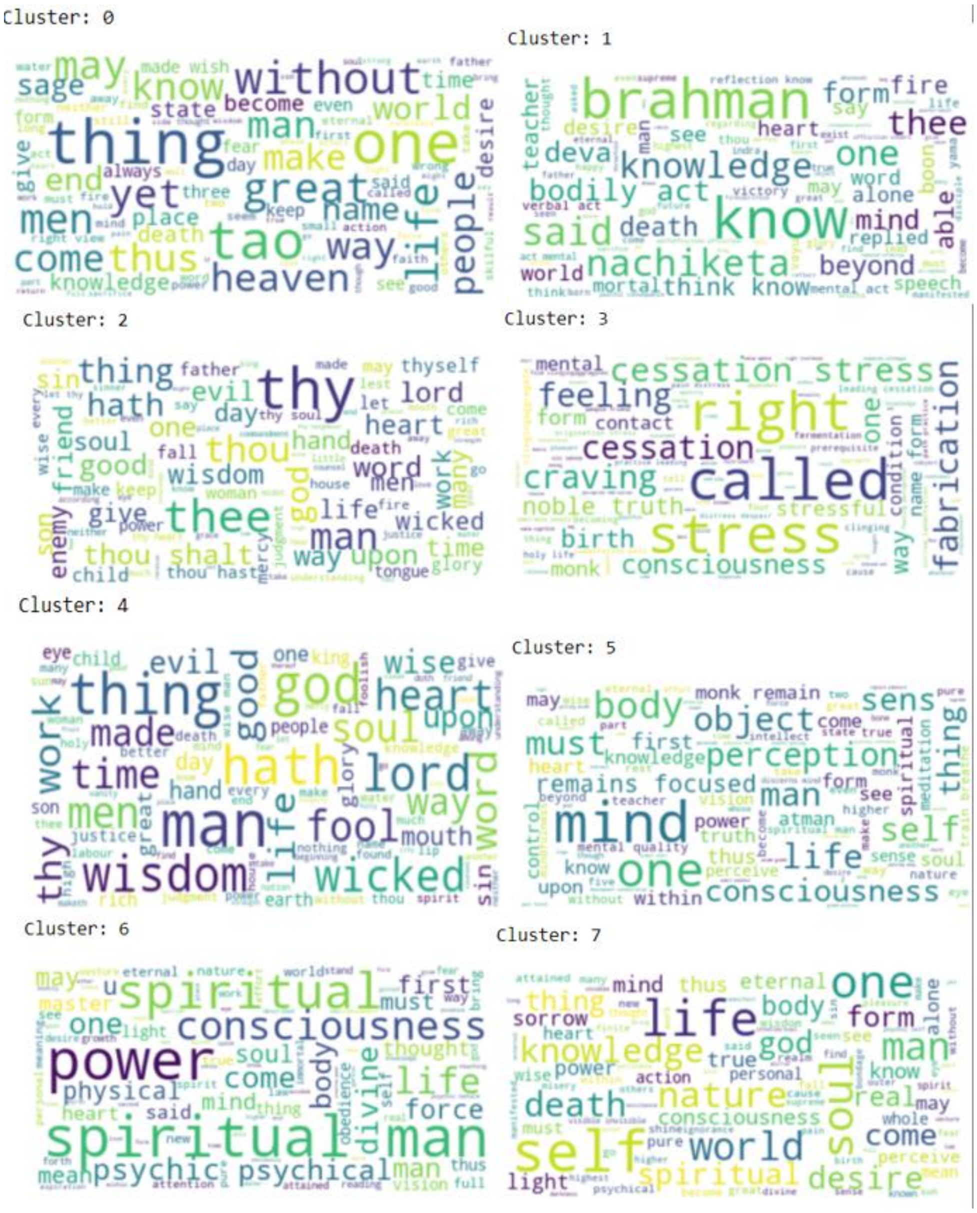
- The clustering models are evaluated using Silhouette score.
- The silhouette coefficient, often known as the silhouette score, is a parameter f or determining how well a classification algorithm is.



K-Means Clustering (TFIDF)	⬆	0.24886
K-Means Clustering (Count Vectorizer)	⬇	0.00421
Spectral Clustering (TFIDF)	⬆	0.21005
Spectral Clustering (Count Vectorizer)	⬇	0.00378
Hierarchical Clustering (TFIDF)	⬈	0.1776
Hierarchical Clustering (Count Vectorizer)	⬇	0.00415

13:Silhouette Score

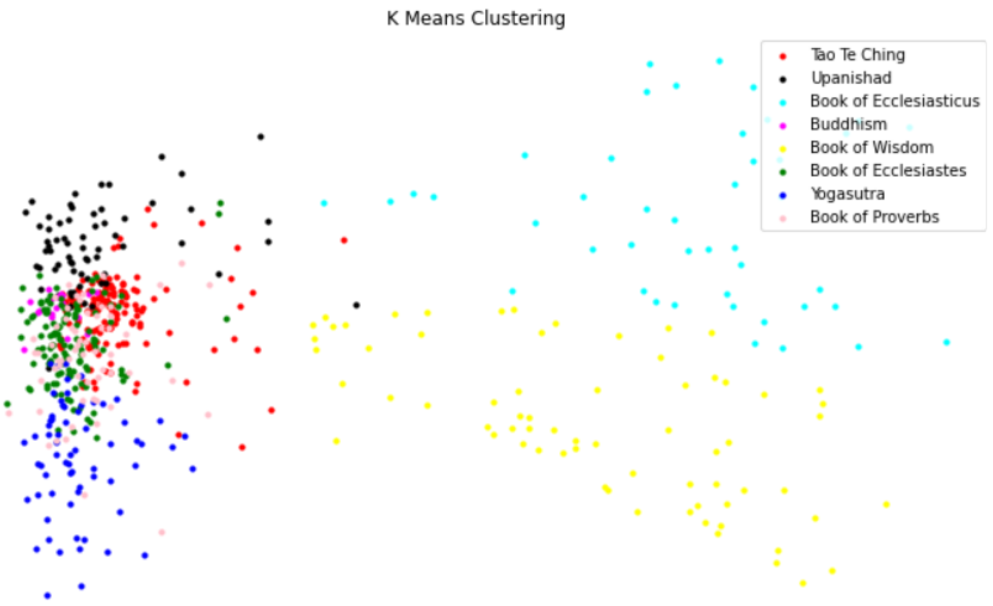
- The silhouette score for K-Means clustering with tfidf vectorized data is the maximum and so we select that model.



14:WordCloud of Clusters

From the word clouds, the following conclusions can be made about each cluster.

Cluster 0	Tao Te Ching
Cluster 1	Upanishad
Cluster 2	Book of Ecclestiasticus
Cluster 3	Buddhism
Cluster 4	Book of Wisdom
Cluster 5	Book of Ecclesiastes
Cluster 6	Yoga sutra
Cluster 7	Book of Proverbs



CONCLUSION

- The given data is pre-processed by removing punctuations, numerical values and converted to lowercase.
- The pre-processed data is then tokenized and lemmatized.
- Two text embedding techniques (TFIDF vectorizer and Count Vectorizer) are used to convert the text data to numerical data for building the clustering models.
- Feature engineering specifically principal component analysis is done to reduce 7581 features to 498 and 281 features respectively.
- Initially traditional methods like elbow plot and silhouette analysis are used to determine the optimal number of Ks. But they do not yield proper number of Ks.
- We then use a new technique called Depth difference method to estimate the optimal number of Ks. The plot shows the optimal number of Ks as 8.
- Three clustering algorithms namely k-means clustering, spectral clustering, and hierarchical clustering are used to cluster the given data.
- Silhouette score is used to select the model.
- We select the k-means clustering model with tfidf data and using the word cloud, label the data.

REFERENCES

https://www.geeksforgeeks.org/python-efficient-text-data-cleaning/

https://towardsdatascience.com/nlp-in-python-data-cleaning-6313a404a470

https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6-unique-ways-perform-tokenization/

https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4#:~:text=Tokenization%20is%20breaking%20the%20raw,the%20sequence%20of%20the%20words.

https://stackoverflow.com/questions/1787110/what-is-the-difference-between-lemmatization-vs-stemming#:~:text=Stemming%20identifies%20the%20common%20root,lemmatized%20as%20%E2%80%9Cgood%E2%80%9D

https://www.geeksforgeeks.org/python-lemmatization-with-nltk/

https://scikit learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

https://www.analyticsvidhya.com/blog/2021/11/how-sklearns-tfidfvectorizer-calculates-tf-idf-values/

https://www.geeksforgeeks.org/using-countvectorizer-to-extracting-features-from-text/#:~:text=CountVectorizer%20is%20a%20great%20tool,occurs%20in%20the%20entire%20text.

https://www.educative.io/edpresso/countvectorizer-in-python

https://www.displayr.com/principal-component-analysis-of-text-data/#:~:text=This%20post%20introduces%20our%20new,of%20loadings%20to%20facilitate%20interpretation.

https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

https://www.researchgate.net/publication/333912095_Estimating_the_Optimal_Number_of_Clusters_k_in_a_Dataset_Using_Data_Depth

https://www.geeksforgeeks.org/ml-spectral-clustering/

https://www.datatechnotes.com/2020/12/spectral-clustering-example-in-python.html

https://www.javatpoint.com/hierarchical-clustering-in-machine-learning

https://www.askpython.com/python/examples/plot-k-means-clusters-python