Effect of Word Embedding Techniques on Clustering of Netflix Movies and TV Shows dataset

Pankaj R. Beldar[1], Rahul Rakhade[2], Vaibhav Khond[3], Prashant Kavale[4], Mugdha Bhadak[5], Milind Bahiram[6]

[1,2,3,4,5,6]*Assistant Professor, K.K.Wagh I.E.E.R.Nashik*

***Abstract*—Netflix is one of the leading over-the-top (OTT) platforms because of its reputation for offering users a wide variety of high-quality streaming movies as well as TV Shows. The reason why Netflix's services are so popular worldwide is that the company uses recent technologies like machine learning, deep learning and Artificial Intelligence to provide consumers with more appropriate and intuitive recommendation. This paper is based on Unsupervised Clustering Analysis on Netflix Movies and TV Shows dataset. Aim of the Project is to form the Clusters based on K mean clustering, Agglomerative Clustering and Affinity Propagation Clustering. We have done Data Preprocessing, Text Cleaning, Exploratory Data Analysis, Vectorization, Implementing Clustering Models, Hyper parameter tuning. Dataset is analyzed with Word2Vec Word Embedding, CounVectorizer and TfidfVectorizer. Out of these Word2Vec has much better performance than other methods. I have Keep Silhouette Score , Elbow Method and Dendrogram as the Selection Criteria for Finding out optimum number of Clusters. We figure out Exploratory Data Analysis, Understanding what type content is available in different countries, Netflix has increasingly focused on TV rather than movies in recent years. Clustering similar content by matching text-based features**

***Index Terms*—clustering, word2vec, countvectorizer, TFIDF vectorizer, NLP, Stemming, and Bag of words**

I. INTRODUCTION

It's fascinating how Netflix applies AI/Data Science/ML to running its operations, such as by implementing algorithms to provide movie recommendations and using AI to guarantee high-quality streaming even at reduced bandwidths. The following are some of the numerous applications of Airdate science, and machine learning at Netflix. Improvement in Netflix’s AI integration has made widespread individualization possible. Simply said, the AI engine keeps an eye on the flow of information and sometimes takes over so that it may make judgments and suggestions at predetermined moments. Netflix's AI considers your viewing habits and hobbies to provide Netflix recommendations. Users can take charge of their multimedia streaming and customize their interactions owing to the system's ability to compile and recommend content based on their preferences.

II. DATA OVERVIEW

There are a total 7787 entities and 12 features in our dataset. About 30.67% data is missing in director, 9.22% in cast, 6.51% in country and 0.0898 % in rating.

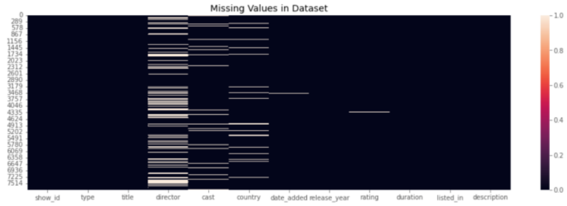


Figure 01: Missing values in dataset

The attribute ‘director’, ’cast’, ‘country’, ‘date added' ,'rating' consists of missing values.

III. DATA CLEANING

The attribute 'director', 'cast', 'country', 'date added', 'rating' consists of missing values. To tackle missing values , we will replace 'country' and 'rating' missing values by the most frequent entity that is 'United States' and 'TV-MA' respectively, Missing values in 'cast' by 'unknown'. There are around 30.68 % values missing in 'director', hence we decide to drop it.

Removal of Missing Values in Dataset
0
289
578
867
1156
1445
1734
2023
2312
2601
2890
3179
3468
3757
4046
4335
4624
4913
5202
5491
5780
6069
6358
6647
6936
7225
7514
show_id
type
title
cast
country
date_added
release_year
rating
duration
listed_in
description
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100

Figure 02: After Removal of Missing Values

No duplicate values exist in the whole dataset.

IV. EXPLORATORY DATA ANALYSIS

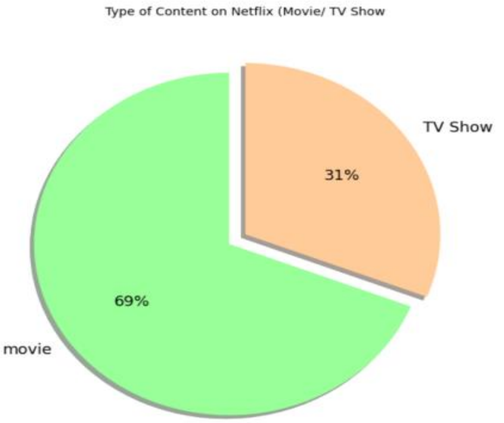


Figure 03: Distribution of Movies and TV Shows

69% of the content available on Netflix are movies; the remaining 31% are TV Shows. Netflix has 5377 movies, which is more than double the quantity of TV shows.

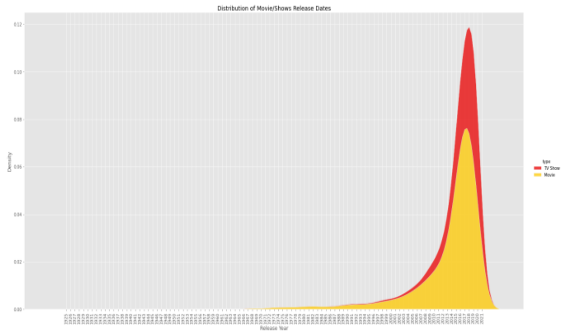


Figure 04: Distribution of Movie/Shows Release year

In recent years more TV Shows are released as compared to Movies on Netflix. Less number of TV shows and Movies were released in 2020-2021 due to corona virus pandemic

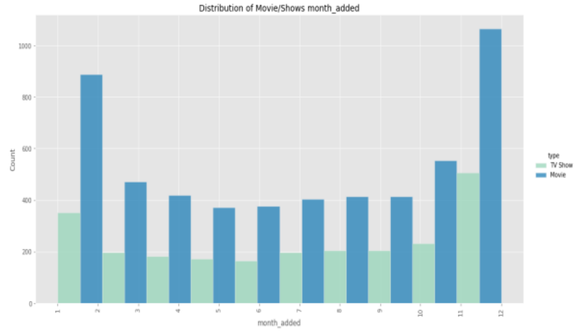


Figure 05: Distribution of Movie/Shows based on month added

Most of the Movies/TV Shows were added in the month of December and January. Number of Movies added on Netflix is more as compared to TV Shows throughout the year.

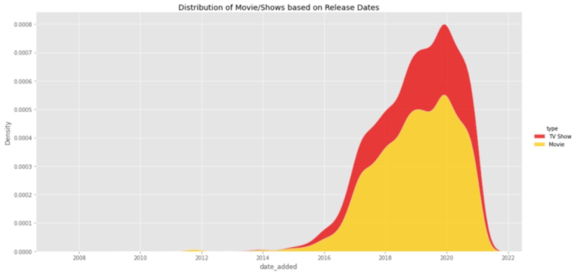


Figure 06: Distribution of Movie/Shows based on date added

In recent few year more number of TV Shows were added on Netflix as compared to Movies , We can say Netflix is more focusing on TV Shows than Movies.

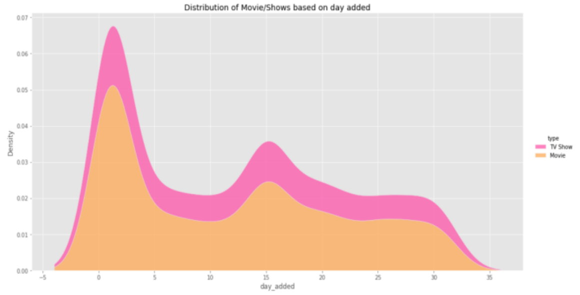


Figure 07: Distribution of Movie/Shows based on day added

Most of the TV Shows/Movies are added on early 5 days of the month, then middle of the month. Again Number of TV Shows added is more as compared to Movies.

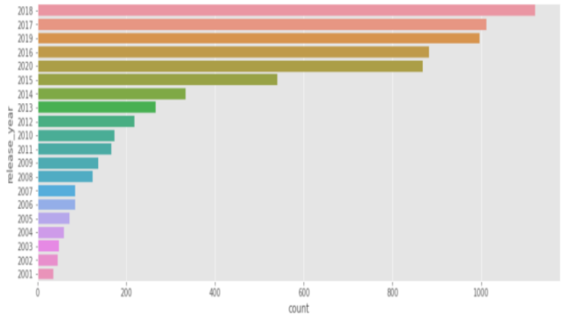


Figure 08: Analyzing how many movies released per year in last 15 years

Number of Movies/TV Shows added on Netflix had drastically increased after 2014. Maximum content was uploaded on Netflix in the year of 2018, after that due to Covid, The count is decreasing slightly.

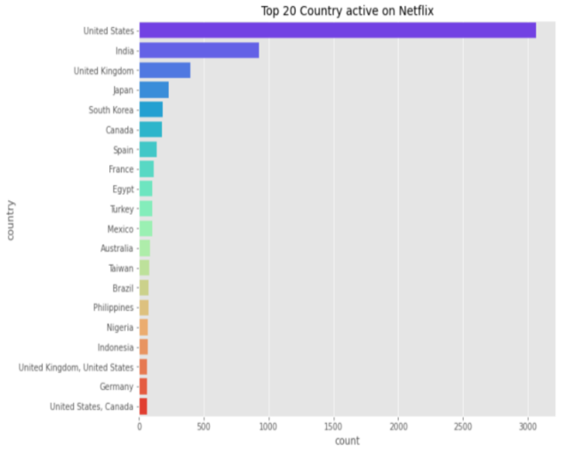


Figure 09: Top 10 Countries that produced content on Netflix

The United States, India, United Kingdom, Japan, South Korea, Canada, Spain, France, Egypt and Turkey are the Top 10 countries which produce most of the content on Netflix.

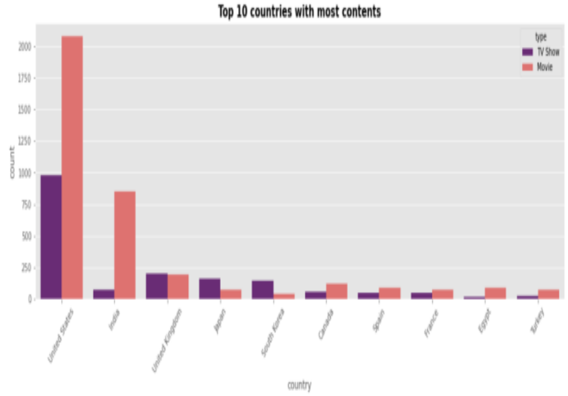


Figure 10: Kind of content is available in different countries in recent years

The United States produced most of the content on Netflix. Also, the number of movies released is more than TV Shows in the United States. In India, Canada, Spain, France, Egypt and Turkey , Most of the content on Netflix is Movies. The United Kingdom has almost equal production of Movies and TV Shows. In Japan and South Korea, Number of TV Shows are available on Netflix. Maturity ratings for each piece of content on Netflix are determined by a local standards organization or by Netflix considering the “frequency and impact of mature content in a TV show or movie”. These ratings are put in place to help your family make informed decisions about the content you are viewing. To get a better understanding of how Netflix breaks down its ratings categories:

1. Little Kids: G, TV-Y, TV-G
2. Older Kids: PG, TV-Y7, TV-Y7-FV, TV-PG
3. Teens: PG-13, TV-14
4. Adults: R, NC-17, TV-MA

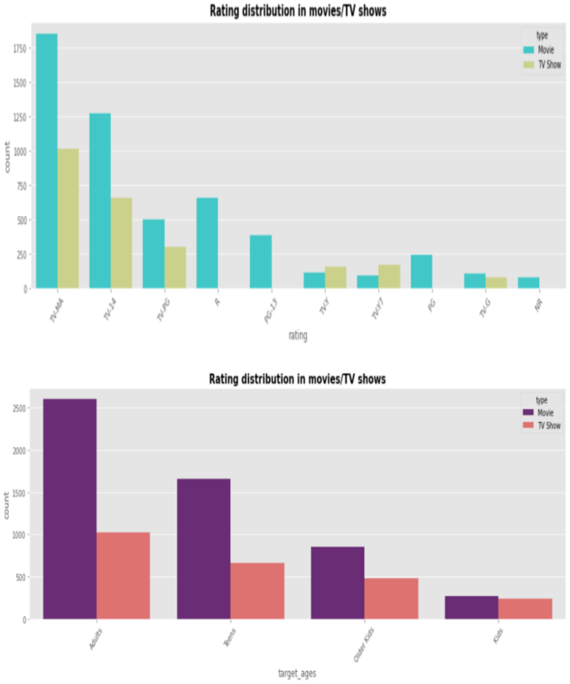


Figure 11: Rating distribution in movies/TV shows

It is observed that, in each category, Quantity of Movies is more than the Quantity of TV Shows. The Availability of the Adult Content is more on Netflix and Least for the Kids.

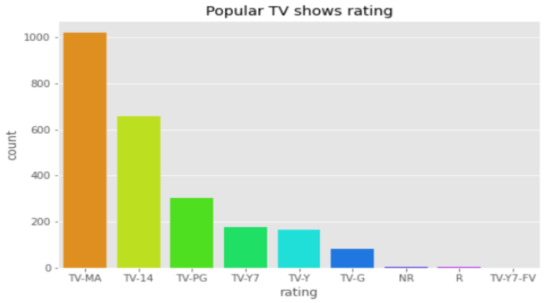


Figure 12: Popular TV shows rating

Popular Movies ratings are TV-MA, TV-14, R, TV-PG, PG-14 and PG. It is observed that Adults and Teens are mostly active on Netflix. Popular TV Shows ratings are TV-MA, TV-14, R, TV-PG, PG-14 and PG. Top 5 Genres in 'TV Shows' are Kid's TV, TV Dramas, TV Crime Shows, TV Comedies, TV Romantic. Top 5 Genres in 'Movies' are Documentaries, Standup Comedy, Dramas and International Movies, Comedies and Independent Movies.

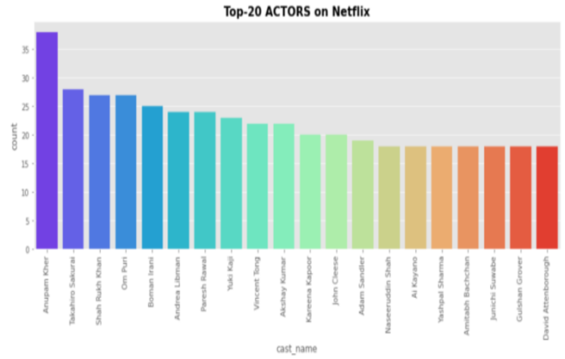


Figure 13: Top 20 actors on Netflix

Famous Actors on Netflix based on the Frequency of their occurrence on screen are Anupam Kher, Takahiro Sakurai, Shah Rukh Khan, Om Puri and Boman Irani and so on.

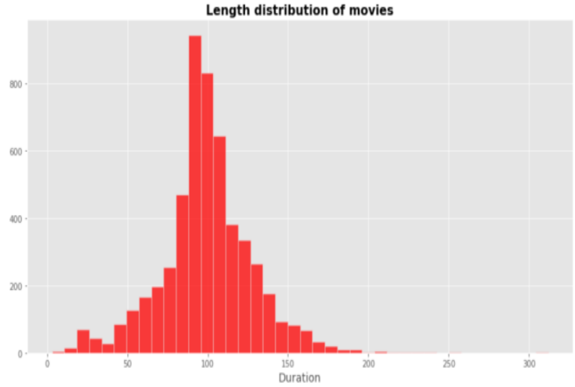


Figure 14: Distribution of Duration of Movies

Most of the Movies/TV Shows have a duration of around 100 min.

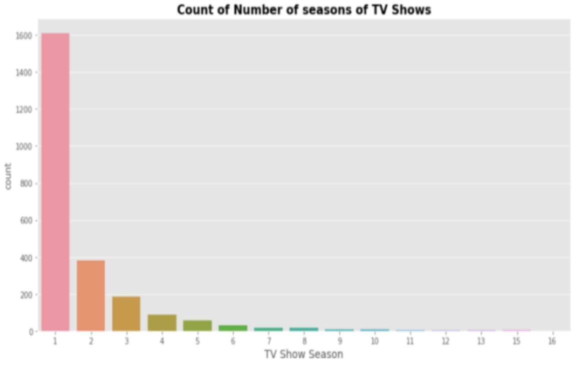


Figure 15: Count of Number of seasons of TV Shows

It is observed that 1608 TV Shows had only one season. The count of longest running TV Shows is very less. The United States produces maximum International TV Shows, TV Dramas, Sci-fi and Fantasy TV shows, International Movies. India, UK, Spain, Egypt, Mexico and Turkey are having most of the Content as Dramas and International Movies.

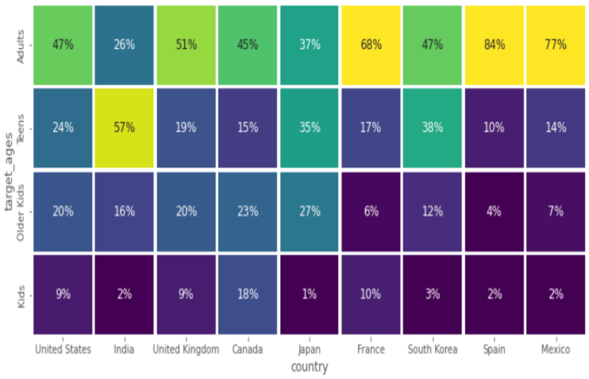


Figure 16: Country wise Content Production in Heatmap

It is observed that content available for kids is less as compared to other categories. Content available for Adults is more in almost every country except India. In India, Most of the content is available for Teens. Netflix should focus on Teen and Adult Contents to generate maximum revenue. Spain and Mexico are producing the highest Adult Content on Netflix almost 84% and 77% respectively.

V. TEXT PROCESSING

1. Removed Punctuations
2. Removed Stop words
3. Removed Short words
4. Convert text to Lower Case
5. Stemming
6. Lemmatizing
7. Tokenizing

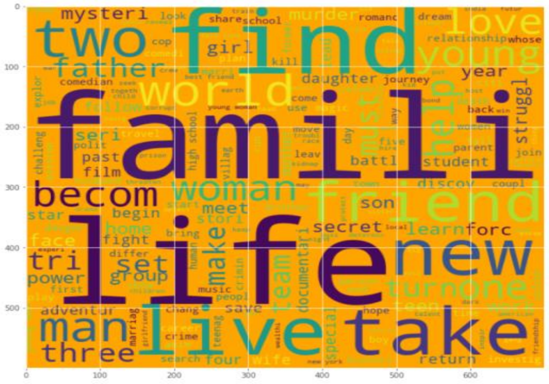


Figure 17: Bag of words

VI. MODEL SELECTION AND HYPER PARAMETER TUNING

1. Feature Selection and Extraction

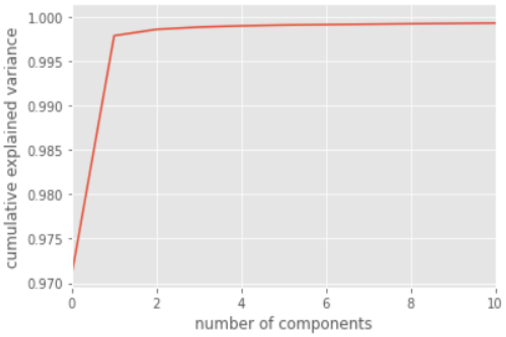
Word Embeddings in NLP is a technique where individual words are represented as real-valued vectors in a lower-dimensional space and captures inter-word semantics. Each word is represented by a real-valued vector with tens or hundreds of dimensions. The word embedding techniques are used to represent words mathematically. One Hot Encoding, TF-IDF, Word2Vec, CountVectorizer, FastText are frequently used Word Embedding methods. One of these techniques (in some cases several) is preferred and used according to the status, size and purpose of processing the data.

PART A: Modeling with Word2Vec for Word Embeddings

1. Word2Vec
Pre-trained vectors trained on a part of the Google News dataset (about 100 billion words). The model contains 300-dimensional vectors for 3 million words and phrases. The phrases were obtained using a simple data-driven approach described in 'Distributed Representations of Words and Phrases and their Compositionality'

2. Encoding Categorical Variables
A one-hot encoding is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

3. Principal component analysis (PCA):
Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

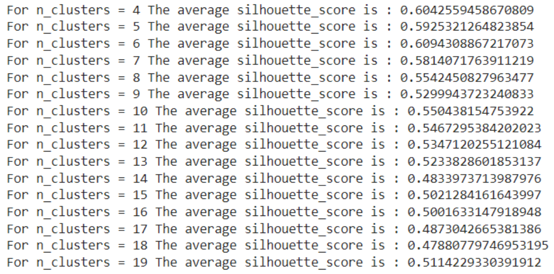


Select 2 PCA Components

4. K-Means Clustering

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar properties. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

5. Calculation of Silhouette score
Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observation belonging to all the clusters: Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The value of the silhouette coefficient is between [-1, 1]. A score of 1 denotes the best meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters. The worst value is -1. Values near 0 denote overlapping clusters.



```
For n_clusters = 4 The average silhouette_score is : 0.6042559458670809
For n_clusters = 5 The average silhouette_score is : 0.5925321264823854
For n_clusters = 6 The average silhouette_score is : 0.6094308867217073
For n_clusters = 7 The average silhouette_score is : 0.5814071763911219
For n_clusters = 8 The average silhouette_score is : 0.5542450827963477
For n_clusters = 9 The average silhouette_score is : 0.5299943723240833
For n_clusters = 10 The average silhouette_score is : 0.550438154753922
For n_clusters = 11 The average silhouette_score is : 0.5467295384202023
For n_clusters = 12 The average silhouette_score is : 0.5347120255121084
For n_clusters = 13 The average silhouette_score is : 0.5233828601853137
For n_clusters = 14 The average silhouette_score is : 0.4833973713987976
For n_clusters = 15 The average silhouette_score is : 0.5021284161643997
For n_clusters = 16 The average silhouette_score is : 0.5001633147918948
For n_clusters = 17 The average silhouette_score is : 0.4873042665381386
For n_clusters = 18 The average silhouette_score is : 0.47880779746953195
For n_clusters = 19 The average silhouette_score is : 0.5114229330391912
```

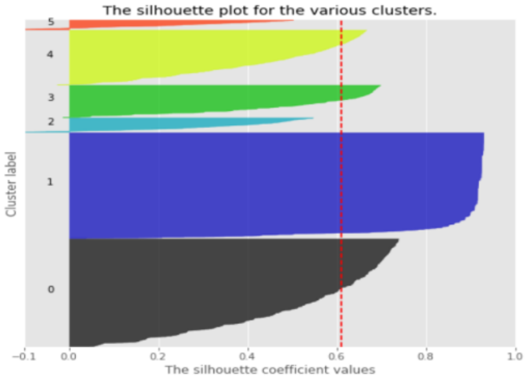


Figure 18: Silhouette score for optimum clusters

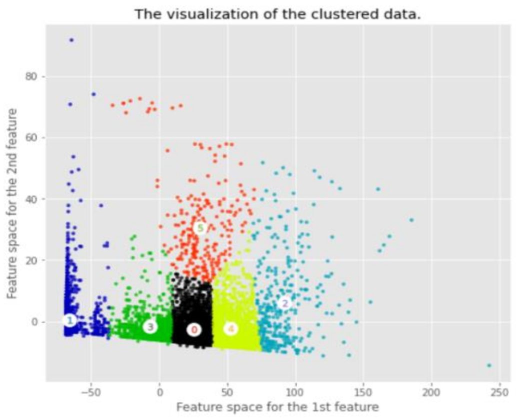


Figure 19: Silhouette score for optimum clusters

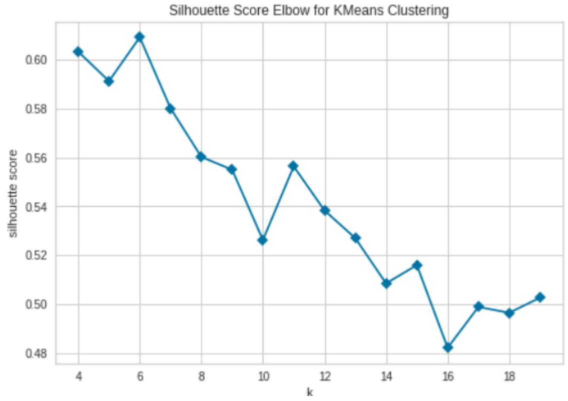


Figure 20: Silhouette score for all clusters

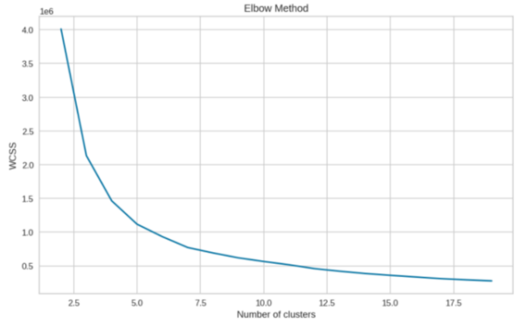
Maximum silhouette score is 0.6 for cluster number 6

Elbow Method to get number of clusters

The K-Elbow Visualizer implements the “elbow” method of selecting the optimal number of clusters for K-means clustering. The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center. When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the “elbow” (the point of inflection on the curve) is the best value of k. The “arm” can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.

6. Elbow Curve Method

Perform K-means clustering with all these different values of K. For each of the K values, we calculate average distances to the centroid across all data points. Plot these points and find the point where the average distance from the centroid falls suddenly (“Elbow”).



Perform Clustering considering k=5.

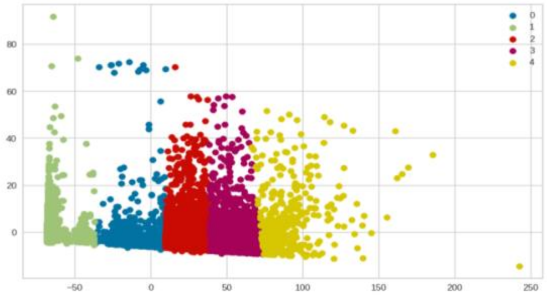
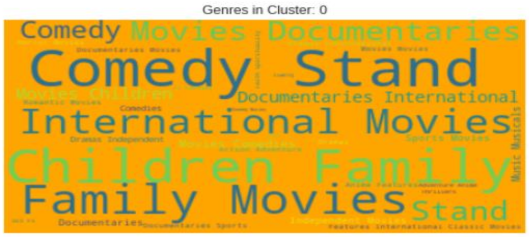
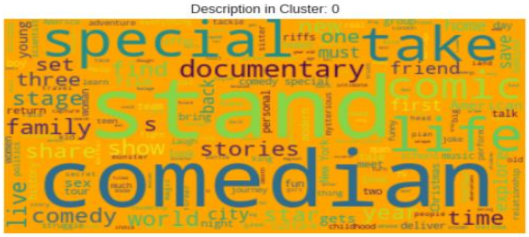


Figure 21: Kmean with 5 clusters





7. Select number of clusters for Agglomerative clustering using Dendrogram

A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The key to interpreting a hierarchical cluster analysis is to look at the point at which any given pair of cards “join together” in the tree diagram. Cards that join together sooner are more similar to each other than those that join together later.

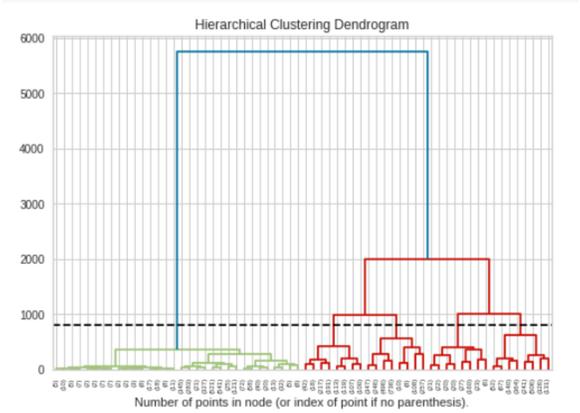
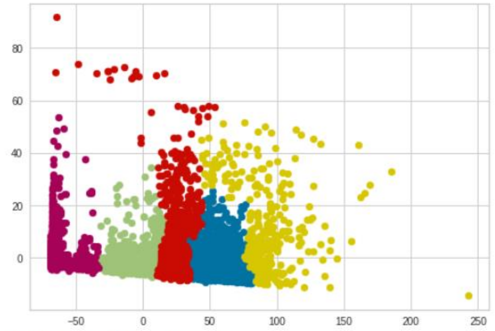


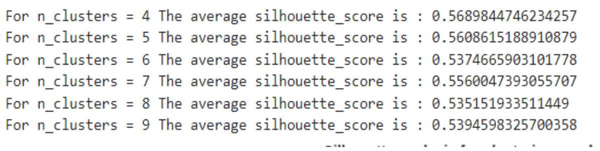
Figure 22: Dendrogram

Number of clusters from Dendogram are 5



```
Silhouette Coefficient: 0.561
davies_bouldin_score 0.683
```

Select number of clusters for Agglomerative clustering using silhouette score



```
For n_clusters = 4 The average silhouette_score is : 0.5689844746234257
For n_clusters = 5 The average silhouette_score is : 0.5608615188910879
For n_clusters = 6 The average silhouette_score is : 0.5374665903101778
For n_clusters = 7 The average silhouette_score is : 0.5560047393055707
For n_clusters = 8 The average silhouette_score is : 0.535151933511449
For n_clusters = 9 The average silhouette_score is : 0.5394598325700358
```

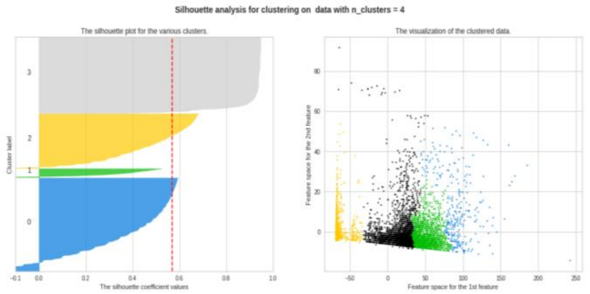
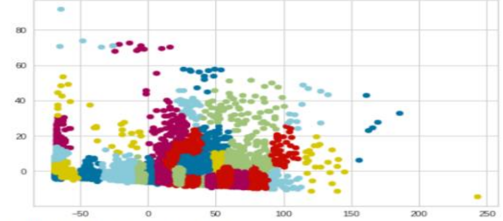


Figure 23: Agglomerative clustering using silhouette score

8. Affinity Propagation Clustering

Affinity Propagation, instead, takes as input measures of similarity between pairs of data points, and simultaneously considers all data points as potential examples. Real-valued messages are exchanged between data points until a high-quality set of examples and corresponding clusters gradually emerges.



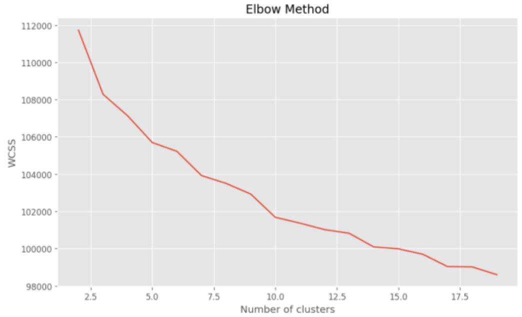
```
Silhouette Coefficient: 0.406
0.7531270201179954
```

Figure 24: Affinity Propagation Clustering using silhouette score

PART B: Modeling with CountVectorizer and TfidfVectorizer

1. K Mean with CountVectorizer and TfidfVectorizer

TF-IDF stands for term frequency-inverse document frequency and it is a measure, used in the fields of information retrieval (IR) and machine learning, that can quantify the importance or relevance of string representations (words, phrases, lemmas, etc) in a document amongst a collection of documents. CountVectorizer means breaking down a sentence or any text into words by performing preprocessing tasks like converting all words to lowercase, thus removing special characters. In NLP models can't understand textual data they only accept numbers, so this textual data needs to be vectorized.



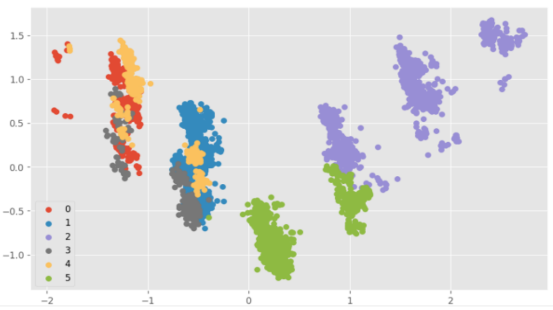
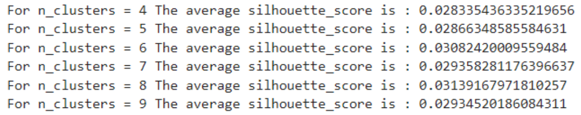


Figure 25: K Mean with CountVectorizer and TfidfVectorizer

2 Agglomerative Clustering with CountVectorizer and TfidfVectorizer



```
For n_clusters = 4 The average silhouette_score is : 0.0283354363352196556
For n_clusters = 5 The average silhouette_score is : 0.02866348585584631
For n_clusters = 6 The average silhouette_score is : 0.03082420009559484
For n_clusters = 7 The average silhouette_score is : 0.02935828117639663 7
For n_clusters = 8 The average silhouette_score is : 0.03139167971810257
For n_clusters = 9 The average silhouette_score is : 0.02934520186084311
```

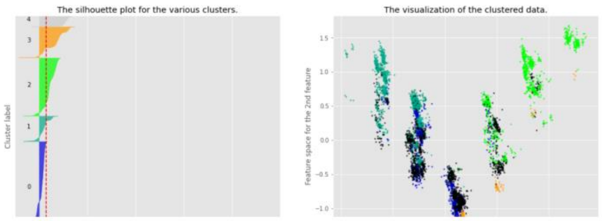
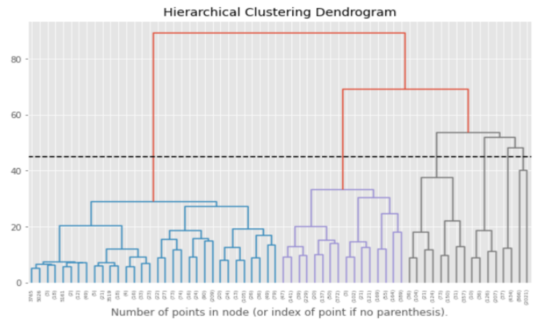
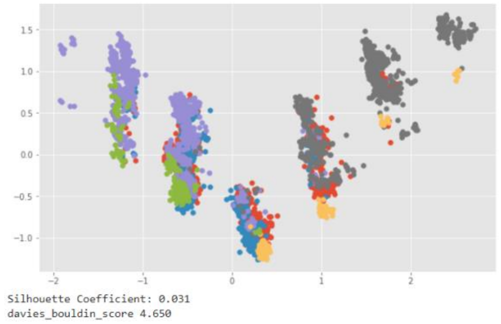


Figure 26: Agglomerative Clustering with CountVectorizer and TfidfVectorizer

3. Agglomerative Clustering with Dendrogram and CountVectorizer



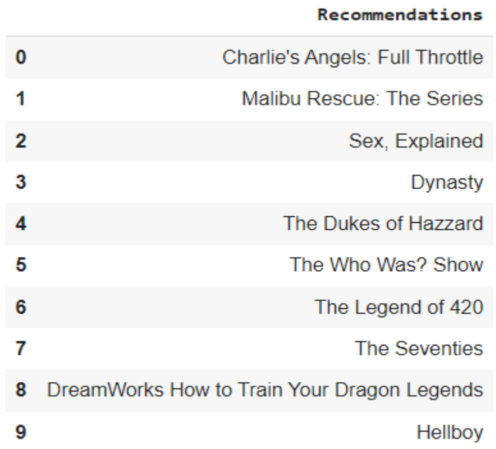


```
Silhouette Coefficient: 0.031
davies_bouldin_score 4.650
```

Figure 26: Agglomerative Clustering with Dendrogram and CountVectorizer

VII. RECOMMENDATION SYSTEM

Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis



	Recommendations
0	Charlie's Angels: Full Throttle
1	Malibu Rescue: The Series
2	Sex, Explained
3	Dynasty
4	The Dukes of Hazzard
5	The Who Was? Show
6	The Legend of 420
7	The Seventies
8	DreamWorks How to Train Your Dragon Legends
9	Hellboy

VIII. SUMMARY

Table no 1: Summary of Models

Sr. no.	Model	Number of Clusters
Part A: Clustering Mdels with word2vec		
1	K-Means with silhouette_score with word2vec	6
2	K-Means with Elbow method with word2vec	5
3	Agglomerative Clustering with dendogram with word2vec	5
4	Agglomerative Clustering with silhouette_score with word2vec	4
5	Affinity propagation clustering with woed2vec	5
Part B: Clustering Mdels with CountVectorizer		
1	K-Means with Elbow method with countvectorizer	6
2	Agglomerative Clustering with dendogram with countvectorizer	6
3	Agglomerative Clustering with silhouette_score with countvectorizer	6

IX. CONCLUSION

I. Exploratory Data Analysis

1. The attribute ‘director’, ‘cast', 'country' ,'date added', 'rating' consists of missing values. To tackle missing values, we will replace 'country' and 'rating' missing values by the most frequent entity that is 'United States' and 'TV-MA' respectively. Missing values in 'cast' by 'unknown'. There are around 30.68 % values missing in 'director', hence we decide to drop it.

69% of the content available on Netflix are movies; the remaining 31% are TV Shows.

2. Netflix has 5377 movies, which is more than double the quantity of TV shows. In recent years more TV Shows are released as compared to Movies on Netflix. Less number of TV shows and Movies were released in 2020-2021 due to the corona virus pandemic. Most of the Movies/TV Shows were added in the month of December and January.
3. Number of Movies added on Netflix is more as compared to TV Shows throughout the year. In recent few year more number of TV Shows were added on Netflix as compared to Movies , We can say Netflix is more focusing on TV Shows than Movies.
4. The United States, India, United Kingdom, Japan, South Korea, Canada, Spain, France, Egypt and Turkey are the Top 10 countries which produce most of the content on Netflix. The United States produced most of the content on Netflix. Also, the number of movies released is more than TV Shows in the United States. In India, Canada, Spain, France, Egypt and Turkey, Most of the content on Netflix is Movies. The United Kingdom has almost equal production of Movies and TV Shows. In Japan and South Korea, Number of TV Shows are available on Netflix.
5. It is observed that, in each category, Quantity of Movies is more than the Quantity of TV Shows.The Availability of the Adult Content is more on Netflix and Least for the Kids.
6. Popular Movies ratings are TV-MA, TV-14, R, TV-PG, PG-14 and PG. It is observed that Adults and Teens are mostly active on Netflix. Popular TV Shows ratings are TV-MA, TV-14, R, TV-PG, PG-14 and PG. Top 5 Genres in 'TV Shows' are Kid's TV, TV Dramas, TV Crime Shows, TV Comedies, TV Romantic. Top 5 Genres in 'Movies' are Documentaries, Stand up Comedy, Dramas and International Movies, Comedies and Independent Movies. It is observed that 1608 TV Shows has only one season. The count of longest running TV Shows is very less.
7. Famous Actors on Netflix based on the Frequency of their occurrence on screen are Anupam Kher, Takahiro Sakurai, Shah Rukh Khan, Om Puri and Boman Irani and so on. Most of the Movies/TV Shows have a duration of around 100 min. The United States produces maximum International TV Shows, TV Dramas, Sci-fi and Fantasy TV shows, International Movies. India, UK, Spain, Egypt, Mexico and Turkey are having most of the Content as Dramas and International Movies.
8. It is observed that content available for kids is less as compared to other categories. Content available for Adults is more in almost every country except India. In India, Most of the content is available for Teens. Netflix should focus on Teen and Adult Contents to generate maximum revenue. Spain and Mexico are producing the highest Adult Content on Netflix almost 84% and 77% respectively.

II. Clustering with Word2vec

1. K-Means with 0.6092 silhouette score with word2vec has an optimum number of clusters as 6.
2. K-Means with Elbow method with word2vec has 5 optimum clusters.
3. Agglomerative Clustering with dendrogram with word2vec has 5 optimum clusters.
4. Agglomerative Clustering with 0.53 silhouette score with word2vec gives 4 clusters.
5. Affinity propagation clustering with word2vec has 5 optimum clusters.

III. Clustering with CountVectorizer

1. It is observed that , after using CountVectorizer and tfidfVectorizer, we get the less silhouette score as 0.032
2. Hence we can say word2vec word embedding method is more suitable for our model.

Winner Model

I have used Principal component Analysis for feature reduction, Recommended System is also designed for getting recommendation of movies and TV Shows. Hyper parameter Tuning is done in every model to get optimum results.

K-Means with word2vec with 6 optimum clusters with 0.6092 silhouette score

Github Depository: https://github.com/pankaj-beldar/Capstone_Project_04_Netflix_Movies_and_TV_Shows_Clustering

X. FUTURE WORK

We can use different Word2vec vectors for further analysis. Some other methods of clustering like Density-based, Distribution-based, Centroid-based, DBSCAN clustering algorithm, Gaussian Mixture Model algorithm, BIRCH algorithm.

REFERENCES

[1] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

[3] https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html

[4] https://help.netflix.com/en/node/2064

[5] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html