Dec 07 2021 Machine Learning Research Paper

Revisiting Genre Classification Through NLP: Topic Modeling and Dissimilarity Measurement with Textual Data

Yeni Hwang, Young Sir Rha

Abstract

Research Summary

In this paper, we reclassify films in the Disney Plus dataset by deploying the widely used NLP method of topic modeling (LDA). Through topic modeling, we are able to utilize unstructured textual data to obtain an in-depth understanding about the films. We utilize the latent topics extracted from topic modeling along with other features from the dataset to conduct K-means clustering to establish a new classification of films, which go beyond the current genre classification by incorporating textual data. In addition, we employ hierarchical clustering to quantify the level of dissimilarity in films in Disney Plus. We introduce a framework of using the height score as a variety index. This index represents the level of dissimilarity of films in a OTT platform. We compare the variety index of Disney Plus and Hulu to show superiority in film variety.

Business Summary

The OTT service industry is a very competitive industry. Many video streaming platforms compete over market share with similar strategies and prices. However, many users are subscribed to multiple competing platforms, a phenomenon referred to as multi-homing, to fill in the gaps between the variety of films that these platforms provide (Stoll)[1]. This phenomenon reflects consumer's pressing need for a platform with a wide range of variety of content. In this paper, we use textual data of films including title, plot, and genre to reclassify films in Disney Plus. Such text-based classification takes into account latent and dimensional aspects of film content, going beyond traditional classification of genre. This allows Disney Plus to understand what kinds of content exactly its products are offering to its users and design its business strategy accordingly. In addition, we establish a novel framework to quantify the level of variety in the platform's offerings and suggest that the quantified level of variety can be used to monitor and manage the products of OTT platforms.

---

[1] Stoll, Julia. "Share of subscription video-on-demand (SVOD) subscribers who also subscribe to other services in the United States as of December 2020, by service." *statista*, Julia Stoll, 5 11 2021, https://www.statista.com/statistics/778912/video-streaming-service-multiple-subscriptions/. Accessed 21 11 2021.

Table of Contents

1. Introduction

OTT stands for “over-the-top” and refers to the providing streaming video content to customers through the internet. It is different from traditional video content services such as television in that customers can choose which film to watch from a wide range of choices that is provided simultaneously. Trending OTT services such as Netflix, Disney Plus, Amazon Prime and Hulu all fall under this industry category. According to Statista Research Department, in 2020, approximately 2.13 billion people were using OTT service platforms worldwide (Statista Research Department)[2]. According to the Advertising & Media Outlook, the number is expected to increase to as much as roughly 2.71 billion people globally by 2025. OTT service is undoubtedly the future of entertainment.

A noticeable consumer behavior in this industry is that many OTT platform users multihome, which means that they use multiple OTT platforms simultaneously. In the United States for example, a survey[3] from Reelgood revealed that Netflix had the most subscribers and that the vast majority of users were multi-homing (Stoll). Multi-homing behavior was consistent in other OTT service platforms such as Disney Plus, Hulu, Amazon Prime, HBO Max, etc. Such statistics show that users are in demand of filling in the gaps between the range of content that different platforms provide. This is an opportunity for platforms to appeal to users by promoting the variety of content they provide. However, there is no uniform measure of the level of variety in the OTT service industry.

In this paper, we utilize topic modeling, an NLP method, with Disney Plus film data and incorporate textual data of the films, such as plot, title, and genre to create a new classification of films. Through topic modeling, a classification process of identifying abstract topics from documents, films will be reclassified into “topics” based on textual data. Using the created topics, we aim to quantify the level of variety of films in Disney Plus. Moving forward, we cluster the topics based on similarity with other additional features from the dataset using K-means clustering in order to establish a more descriptive classification model based on textual description of the films. Detailed description of the methodology used is provided in section 3, Methodology. The result of our analysis has a greater business impact on Disney Plus in that it allows the company to understand what exactly its products are about. Our analysis results also provide a symbolic number to represent the level of film variety, which is the core of consumer needs in the OTT industry.

---

[2] Statista Research Department. “Number of OTT video users worldwide from 2017 to 2025.” *statista*, 5 7 2021, https://www.statista.com/forecasts/1207843/ott-video-users-worldwide. Accessed 21 11 2021.

[3] Stoll, Julia. “Share of subscription video-on-demand (SVOD) subscribers who also subscribe to other services in the United States as of December 2020, by service.” *statista*, Julia Stoll, 5 11 2021, https://www.statista.com/statistics/778912/video-streaming-service-multiple-subscriptions/. Accessed 21 11 2021.

2. Current Film Classification Methods

Current classification of films is based on a traditional framework that is widely known as genre. Disney Plus has 19 categories in Movie Genres and 18 categories in TV Shows genres, grouped based on the traditional framework of classification including documentaries, animated movies, live actions, etc. Other OTT platforms such as Amazon Prime Videos and Hulu also show a similar pattern. Netflix, the strongest player in the OTT platform industry, has hired humans to assign tags to all of Netflix TV shows and movies to create hyper specific micro genres, such as "Visually-striking nostalgic dramas" or "Understated romantic road trip movies" (Blattmann)[4]. Such micro classification led to 3,000 unique categories on Netflix (Moore)[5]. These dozens of genres aim to describe the characteristics of films. However, there is no uniform standard of genre and it varies across the globe. Moreover, many modern films are not confined to one genre. Therefore, it has become a very confusing task to classify films under the traditional and theoretical classification of genre. Nonetheless, classifying films effectively is a challenge to OTT service platforms, as classification of films is a fundamental factor in product management in that it entails understanding the content and audience of films. In this paper, we introduce a machine learning approach to classify films of Disney Plus to obtain a deeper insight into the content they deliver and to examine how diverse they are.

3. Methodology

To account for the limitations of the current classification by genre, we suggest a complementary classification system by employing text-based classification. In this section, we will first give a brief description of topic modeling, a machine learning tool that allows us to make use of latent information by mining text data. Then we will explain how we utilized topic modeling results to create an alternative, text-based clusters of films using K-means clustering. Lastly, we will leverage hierarchical clustering to quantify the dissimilarities between clusters of disney that heights of cluster imply.

3.1 Topic Modeling

Topic modeling is a method of Natural Language Processing. Thanks to advancements in modern big data technology, we are now able to conduct quantitative analysis with unstructured textual data. This means that human words

---

[4] Blattmann, Josefina. "Netflix: Binging on the Algorithm | by Josefina Blattmann." *UX Planet*, 2 August 2018, https://uxplanet.org/netflix-binging-on-the-algorithm-a3a74a6c1f59. Accessed 22 November 2021.

[5] Moore, Kasey. "The Netflix ID Bible – Every Category on Netflix in 2021." *What's on Netflix*, 29 September 2021, https://www.whats-on-netflix.com/news/the-netflix-id-bible-every-category-on-netflix/. Accessed 22 November 2021.

can also be fed into machine learning algorithms for quantitative analysis. With topic modeling, we are able to extract latent themes, referred to as topics, from piles of words. The topic modeling assumes that a document (an observation in topic modeling, which is a piece of text, either large or small) consists of a set of topics, and each topic consists of a set of words. In other words, a document is a collection of words that are associated with certain topics, with the words associated with the "heavier" topics in the document being more frequent, and vice versa. A topic modeling algorithm infers the probability distribution of keywords across topics and the distribution of topics across documents by analyzing the pattern of word occurrence in the corpus (Choi, Menon, & Tabakovic, 2021)[6].

3.2 K-means Clustering

K-means is an exploratory data analysis technique used to get an intuition about the structure of the data (Dabbura)[7]. It clusters data into subgroups so that data in each cluster are homogeneous while data in different clusters are heterogeneous. To choose the number of clusters, k, we employed Bayesian information criterion (BIC) as our regularization technique using R software. The algorithm solves k-Means for each k=1,2,3,4,...,n and computes the value below.

$$\min_{k} \; Deviance\ of\ using\ k\ centers + \lambda \{\#\ \text{of clusters}\} \times \{\#\ \text{of features}\}$$

The goal is to find a balance between two variables: the number of clusters (k) and the average variance of the clusters so that the model is not too complex while keeping the average variance minimal. K-means clustering is very popularly used for analyzing the given dataset as it creates latent classifications which have not been explicitly labeled in the raw data.

In this paper, we incorporate the results from topic modeling into clustering the films of Disney Plus. Along with existing data from the original dataset of Disney Plus, we plug in the topic distribution information within each document along with other features such from the dataset to K-means clustering and use the clusters to reclassify its films. With the topic distribution information incorporated in clustering, we are able to cluster films based on latent textual information which gives us a more detailed understanding of film content. As K-means clustering is an unsupervised machine learning tool, the algorithm is immune to human judgements and biases. Therefore, our machine learning approach to reclassifying film categories will provide an unbiased classification that also takes into account descriptive information about the film in the textual data.

---

[6] Choi, Jaeho, et al. "Using machine learning to revisit the diversification–performance relationship." *Wiley*, 2021.

[7] Dabbura, Imad. "K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks." *towards data science*, 17 9 2018, https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. Accessed 21 11 2021.

3.3 Hierarchical Clustering

Hierarchical clustering is another unsupervised machine learning approach for clustering unlabeled datasets. Hierarchical clustering is different from K-means clustering in that it doesn't require a fixed number of clusters and is possible to find appropriate the number by interpreting the dendrogram. Hierarchical clustering starts with treating each data point as an independent cluster. It then follows the following two steps. First, it identifies the two clusters that are closest together. Second, it repeats the merging process of the adjacent two clusters until all clusters are merged into one big cluster. While there are many ways to calculate the distance between two clusters, in this paper we employ the euclidean distance in which the distance between two clusters are calculated based on the length of the straight line between the two. The output of this process is dendrogram, which shows the hierarchical relationship between the clusters. Each observation is placed in the x axis, while the y axis represents the dissimilarity score between two clusters, referred to as height. The dissimilarity score is the height at which any two clusters are joined together. The larger the height is, the more dissimilar the two clusters are in terms of longitude and latitude. The height indicates the order that the clusters were joined, which is why this clustering method is called hierarchical clustering. With hierarchical clustering, we can structure clusters based on data similarities.

4. Data

In this paper, we use a Disney Plus film dataset from Kaggle[8] (Fontes). Each observation is a film that is listed in Disney Plus platform and there are 992 observations in this dataset. There are 19 features for each film including: database ID ("imdb_id"), "title", "plot", "type", film rating ("rated"), the year of release ("year"), exact date of release ("released_at"), the date of addition on Disney Plus ("added_at"), "runtime", "genre", "director", "writer", "actors", "language", countries that a film is provided in "country", "awards", "metascore", imdb user rating ("imdb_rating"), and "imdb_votes". Textual data such as "title", "plot", "genre", "director", "writer", and "actor" will be used in text mining. Feature "plot" consists of a few sentences and provides an overview of the film. Features such as "language" and "country" are categorical data, while "awards" is a string variable with various types of wins and nominations listed.

4.1 Data Cleaning

First, we dropped some columns. Feature "year" was dropped as it was repeated in "released_at". We also dropped "metascore" because there were over 600 null values and dropped "added_at" because over 700 of them were recorded on one data, which may skew the result.

---

[8] Fontes, Raphael. 2020. *Kaggle*, Raphael Fontes, https://www.kaggle.com/unanimad/disney-plus-shows.

Next, we eliminated rows that contained null values in “title”. 90 rows were deleted in this process, leaving us with 894 rows of observations. This was necessary as later on in our topic modeling, features with missing titles would not be able to be linked back to its accountable film.

Then, we manipulated all null values into “NA”, except for those in columns “rated” and “award”. The reason why we transformed the null values in some columns into “NA” was because they were only a few of them, meaning that they won’t have a significant influence on the outcome of our analysis. Turning them into “NA” would allow us to omit these values later in our analysis. We then made 152 “NA”s in “rated” into a factor. We made this decision based on the idea that missing occurrences of these values itself actually has a meaning. Therefore, we made the missing values in these two columns a different level factor.

In “runtime”, there were mins after every number, so we eliminated the string “min”, named it “runtime_min” and added it into the dataset. We then dropped the original “runtime” column from the dataset.

There were missing values in the feature “type”, which were eliminated when we got rid of files with missing titles. We assigned “type” into a level factor of episode, movie, and series.

As the value of “released_at” comes from how old a film is compared to other films, we took a difference of dates in “released_at” and the release date of the latest film in the dataset. The resulting date was changed to a numeric data which was divided by 365.25, the average number of days in a year, and finally multiplied by 12 to get the result in terms of months. We named it “months_from_latest”. This gave us a relative age of films.

Features “language” and “country” were string data that listed one to several kinds of languages and countries. These features were skewed in that English and the United States appeared in the vast majority of film observations. Although it is standard practice to leave out highly skewed data, we altered “language” into dummy variables. This is because language is an important determinant of film classification due to the nature of media comprehensiveness. Leaving out “country” but making “language” into dummy variables was our scheme to incorporate them in clustering the films.

Feature “awards” was a string. It contained information about wins and nominations in two categories: major award and the rest. We extracted only the numerical data from “awards” and classified them under two new features: “Nominates” and “wins” .

Lastly, to prepare textual data for topic modeling, we combined “title”, “plot”, and “genre” into one character feature and named it “film”. We chose these features as our object of topic modeling because these features best represent the content of films. To prepare for textual data analysis, we followed the common practice of text processing by making them all lowercase, removing punctuation and unhelpful words(stopwords) and stemming. Doing so enables us to get meaningful words and their frequency in each document which will transfer into proportion.

4.2 EDA

In the Disney Plus dataset, films are categorized into three types: episode, movie, and series. Movies took up 76% of films in this dataset. The distribution is shown in Figure 1.

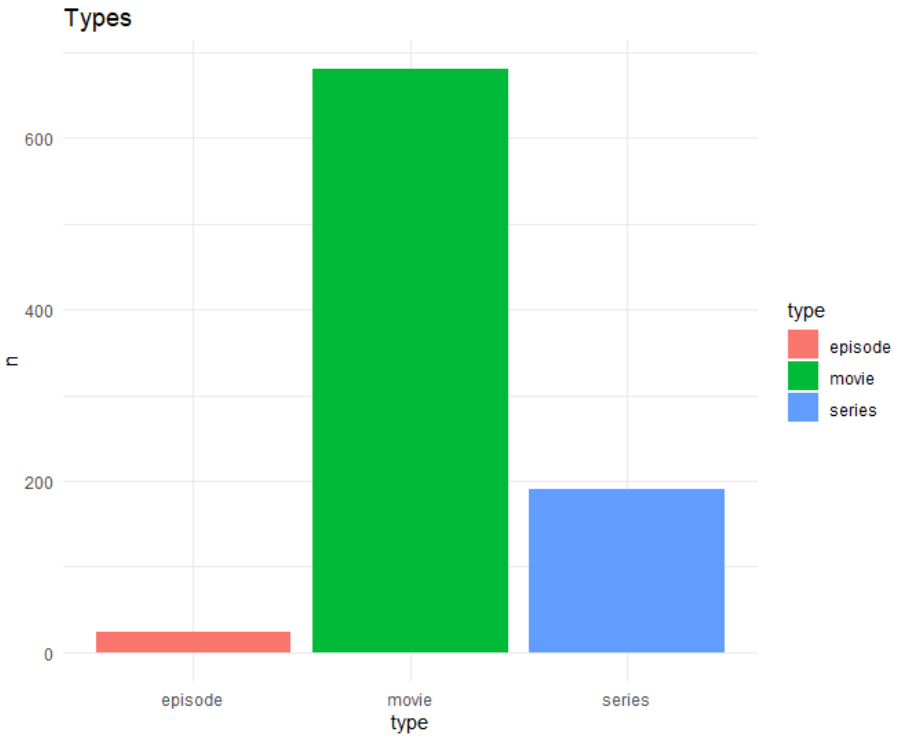

Figure 1

There are 22 unique genres in the Disney Plus dataset. Figure 2 below shows the distribution of each genre across films in the dataset. The top three genres were Family, Comedy, and Genre. From this figure, we can see that many films are categorized under multiple genres. Infact, out of 894 films, there were only 89 films that were categorized under only one genre which means that a vast majority of films spread across genres. This is proof that the traditional categorization under genre itself cannot represent the content of a film.

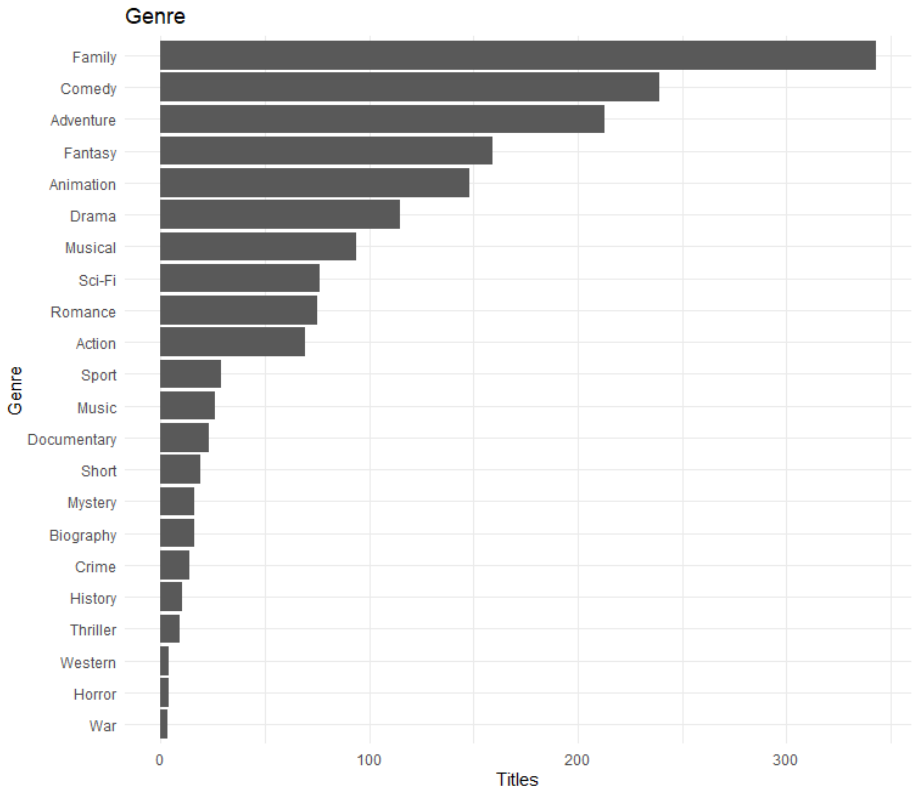

Figure 2

With a motive to see the distribution and density of words in our textual data, we created a word cloud for the whole dataset. But first, from our newly created character “film”, we grouped words into groups, making the unit of analysis bigrams. N-grams of texts are popularly employed in text mining and natural language processing tasks. By using a bigrams, which is a set of adjacent words, we can obtain spatial information on word ordering in the original text data. The experimental results suggest that the bigrams can substantially raise the quality of feature sets (Tan, Wang & Lee, 2002)[9]. Figure 3 is the bigram word cloud of the whole dataset. The more a word appears in the dataset, the bigger the word appears in the word cloud.

[9] Tan, Chade-Meng, et al. “The use of bigrams to enhance text categorization.” *Information Processing & Management*, 2002, p. 529. *ScienceDirect*, https://www.sciencedirect.com/science/article/pii/S0306457301000450.

Figure 3

5. Modeling and Findings

In this section, we describe the modeling and text mining process in detail and explain how the results have a business impact on Disney Plus.

5.1 Topic Modeling: Disney Plus

As explained in the methodology section, we extracted character values from “title”, “plot” and “genre” and for each film combined them into a single document named it “films”. We then turned data into a corpus with all the “films” observations together. We conducted standard text processing including lowercasing, removing punctuation and unhelpful words(stopwords) and stemming in order to format the textual data into a predictable and analyzable form for NLP. We then created a matrix with each document, referred to as “films” in our analysis, in rows and each word in columns. The matrix shows the actual counts of a word in each document. The number of topics is chosen based on the Bayesian information criterion (BIC). Within a vector of assigned topic sizes, it chooses the biggest Bayes factor. Given a selected vector of (5, 10, 15), the algorithm chose five topics. We now have five topics, extracted from the film title, plot and genre descriptions, which broadly capture the various latent aspects of films.

We will first look at the distribution of words in topics. Topic modeling assumes that each topic is made up of distribution of words, which the theta value represents. With theta values, we created a word cloud whose frequency is over 0.02 for each five topics. It is shown in Figure 4. Because some words, such as family, appeared frequently across all topics, we have faded it out to make other words stand out.

For instance, while Topic 5 concerns nature documentaries, based on keywords such as “ocean”, “natur”, and“adventure”, Topic 1 concerns fictional stories, based on the keywords: “fantasi”, “anim”, and “santa”. Also, some of the topics were related to Disney animations: Topic 3 included “donald”, “pluto”, and ”mickey” and Topic 4? included “mermaid” and “aladdin”. We can also infer the age

range of characters of films in some topics, as the plot descriptions share information about characters in the film. Topic 4 has keywords such as “halloweentown”, “boy”, “girl” and “magic”, which implies that the characters in these films are comparably younger than those in Topic 2, which has keywords such as “teenag”, “high” and “student”, implying that the characters are comparably older. As such, the topics excavate and uncover latent features of films.

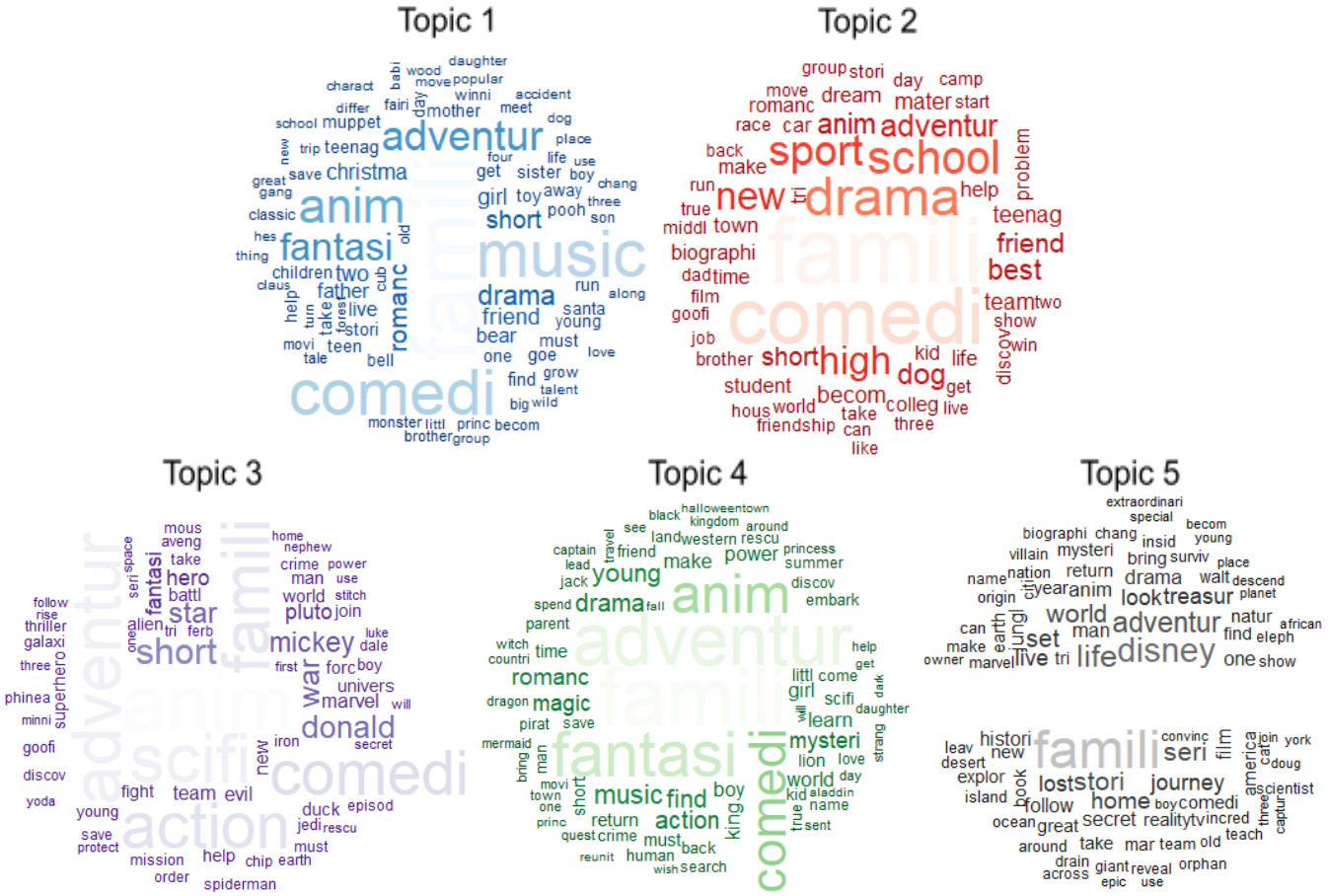

Figure 4

Furthermore, we can look at the relative use of words across topics. Figure 5 shows the top 10 words for each topic. Unlike Figure 4, these 10 words are chosen based on their relative rareness. In other words, these words are unique to the topic in that they are rare in other topics, but have a high in-topic probability.

Topic 1	abl, fiddl, industri, overhear, voic, bricklay, escort, granni, crowley, stewart
Topic 2	1900, 1913, 20yearold, englishman, franci, ouimet, vardon, lilli, moscovitz, sixteenth
Topic 3	council, count, dooku, iii, defeat, defect, milit, stormtroop, vii, allstar

Topic 4	fifteenfoot, gorilla, gregg, jill, ohara, strasser, vengeanc, chines, eighti, valet
Topic 5	brian, earl, environmentalist, hotspot, kennedi, maxwel, oceanograph, photograph, skerri, sylvia

Figure 5

Next, we will look at the topic distribution in documents. Topic modeling assumes that each document also consists of a distribution of topics, which the omega value represents. The document-topic matrix represents the distribution of topics within each document. We can compare the distribution of topics in one document to that of the different documents. The significant value of our classification framework which incorporates textual data comes from the fact that it allows us to distinguish two films that are currently classified under the same genre(s). We will demonstrate this with an example from our analysis. In the Disney Plus dataset, “10 Things I Hate about You” and “Stargirl” are classified under Comedy, Drama, and Romance. However, topic modeling revealed that in fact, they differed significantly in their topic distributions. Figure 6 shows the topic distributions for “10 Things I Hate about You” and “Stargirl”. From this figure, we can see that “10 Things I Hate about You” was heavily associated with topic 1 (0.92), while “Stargirl” was heavily associated with topic 4 (0.71). The reason why they were reclassified as such stems from the fact that descriptive textual data about the films were taken into account in topic modeling. The most obvious difference between the two films can be found in plots of the films. The plot of “10 Things I Hate about You” had the word “teenager”, “popular” which is shown in figure 4, while “Stargirl” had the word “mysterious” and “boy”, “girl” as well. Such finding encourages Disney Plus to revisit their current classification framework.

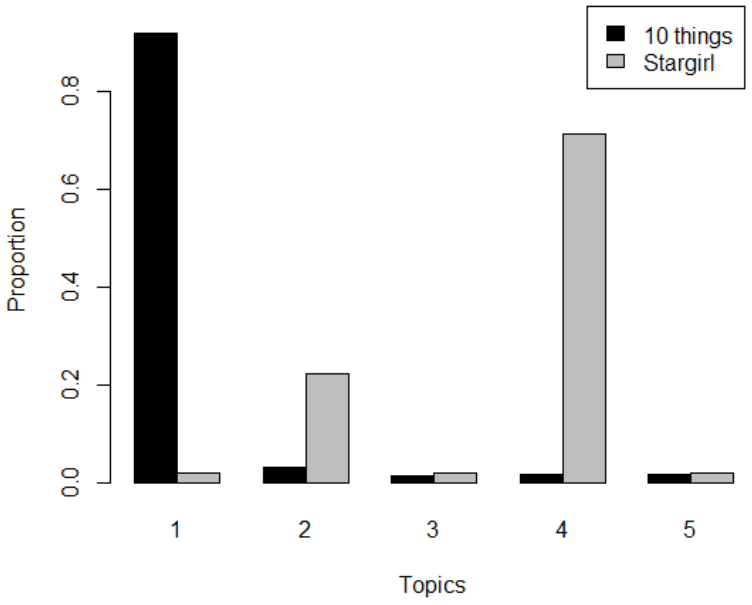

Figure 6

5.2 K-means Clustering with textual data

Furthermore, we utilized the latent topics extracted from topic modeling in K-means clustering to reclassify Disney Plus films based on its homogeneity. We plugged in the topic distribution data, the omega value presented with five topic columns, into K-means clustering algorithm along with other remaining features from the Disney Plus dataset which are: "type", "rated", "runtime", "imdb_rating", "imdb_votes", "months_from_latest", "Nominate", "wins", and all dummies for languages. Five clusters were chosen based on BIC optimization. As a result, we obtained five clusters of Disney Plus films. Figure 7 shows the size of each cluster.

Cluster	1	2	3	4	5
Documents	22	141	349	132	174

Figure 7

By comparing the average (scaled) measures of each feature across topics, referred to as cluster centers, we can have an understanding about the characteristics of each cluster. For example, cluster 5 had a significantly high imdb rating (1.40) compared to other clusters(-0.42, 0.17, 0.03, 0.57). Cluster 5 also had a noticeably high number of award wins (5.26) compared to other clusters (-0.18, -0.11, 0.06, -0.22). Cluster 5 had a relatively high measure of Topic 4 (0.31) compared to other clusters (0.23, 0.13, 0.23, 0.13). A complete view of cluster centers is provided in the exhibit.

5.3 Quantifying dissimilarity

Our framework of quantifying dissimilarity in Disney Plus films employs a measure of distance, height score, from hierarchical clustering, and assigns the score as an index of ‘variety’ in films. While hierarchical clustering is a method to cluster observations based on similarity, it also gives a heights score for each number of clusters, which is the measure of distance between clusters.

In order to cluster Disney Plus films using hierarchical clustering, we inserted the topic distribution data (omega value) extracted using title, plot, and genre data into a hierarchical clustering algorithm. The height score represents the variety in films very well because it was calculated based on descriptive textual information about what the films deliver to the audience.

Figure 8 is a dendrogram of Disney Plus. In the dendrogram, each observation is placed in the x-axis, while the y-axis is the value of the distance metric, referred to as heights, between clusters. Our analysis showed that the textual data films in the Disney Plus had a maximum height score of 1.34 when it was divided into 2 big clusters.

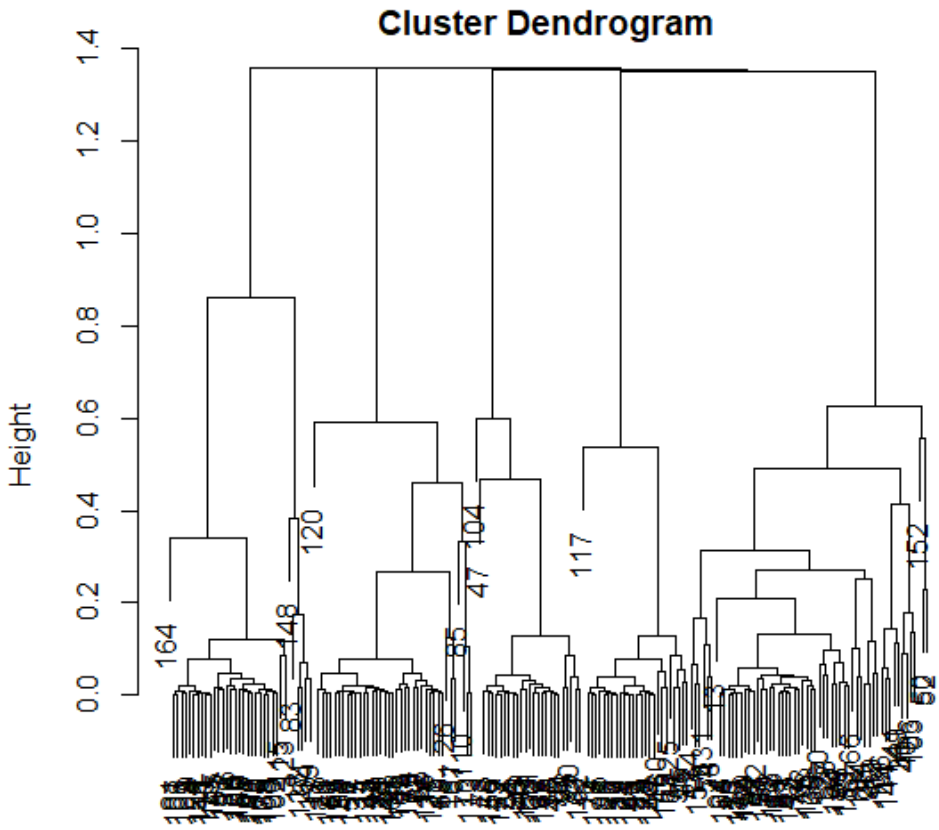

Figure 8

Our framework has great value in that it quantifies the level of variety of products that Disney Plus offers. Without the variety index, measuring the level of product variety has to rely on other factors such as the number of films and the number of genres, which are not directly related to the level of variety because it fails to consider what the films truly deliver.

The variety index becomes more insightful when it is compared across similar businesses. To illustrate, we calculated the variety score of series on Hulu, another big player in the OTT industry which was acquired by Disney in 2019. We worked with a dataset from Kaggle (Alaham)[10], which contained information including title ("show.name"), plot ("show.description"), genre ("show.genre") of 109 Hulu films. To compare the results of text analysis of Disney Plus films to that of Hulu, from the Hulu dataset, we followed the same analysis steps with the same features that we chose for Disney Plus: title, plot and genre. After processing the text, we assigned a vector of (5, 10, 15) to the topic modeling algorithm and it chose five topics. We put a five topics distribution matrix into hcluster model. The dendrogram of Hulu is shown in Figure 9. Films in the Hulu dataset had a max height score of 1.39 approximately, when it was divided into two big clusters.

[10] Alaham, Sanjana. *HULU Shows*. 2021. *Kaggle*, https://www.kaggle.com/sanjanaalaham/hulu-shows/version/1?select=HuluRaw.csv. Accessed 2021.

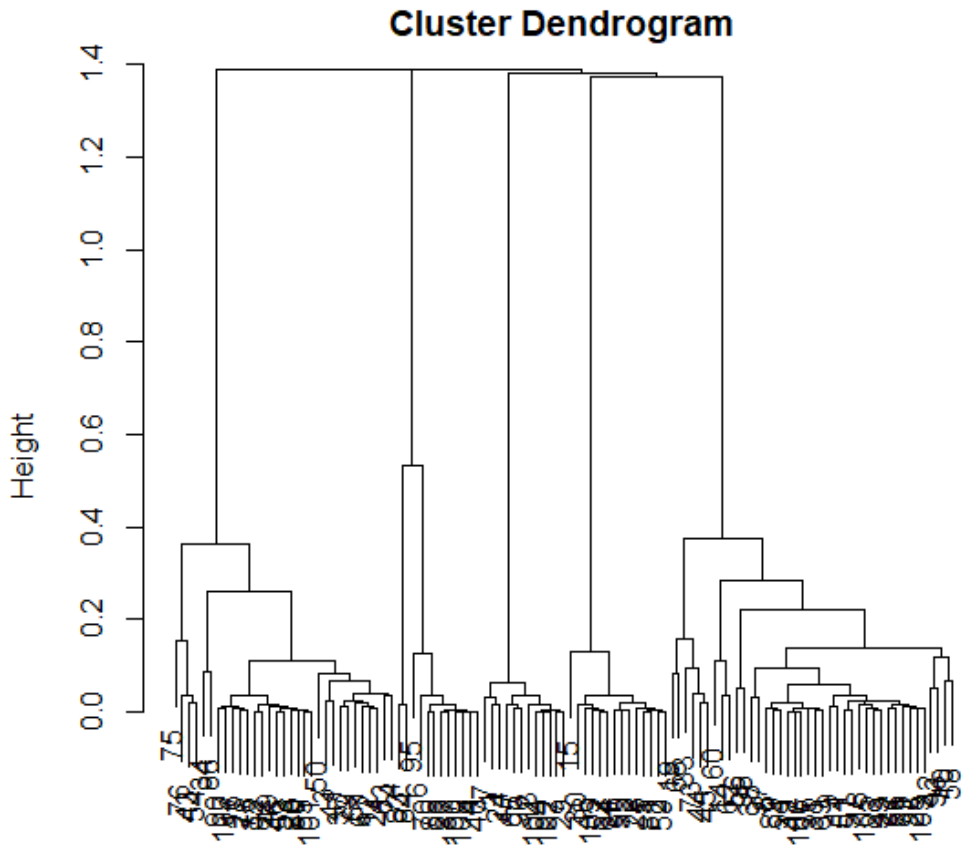

Figure 9

As films in the Hulu dataset only contained TV Shows, we conducted hierarchical analysis on Disney Plus only with its series/shows to conduct a fair analysis. Both textual data from Disney Plus and Hulu had a similar number of words, each having 3,701 and 3,768 words respectively. Disney Plus series/shows had a max height score of 1.35. Comparing the two platform's height scores on 1 to 5 clusters, Disney Plus had a lower height score except for when there were five clusters. In other words, distances between the clusters in Disney Plus data tend to be longer and less similar to each other than Hulu's. The results are shown in Figure 10.

Number of clusters	2	3	4	5	6
Disney Plus	1.3589	1.3560	1.3523	1.3483	0.8610
Hulu	1.3903	1.3871	1.3813	1.3746	0.5345

Figure 10

6. Business impact and Conclusion

This paper introduces a framework to utilize NLP into strategy. Our framework of incorporating textual data into reclassifying products and creating new measures can be applied to any businesses and industries. Textual data provides significant insight that was unobtainable only with structured data. In this paper, we used NLP techniques to convert textual data into quantitative data for analysis. By applying topic modeling to the corpus of Disney Plus film data containing textual information, we were able to extract latent topics and see the distribution of words within topics as well as the distribution of topics within documents. Incorporating such text-mined information into classifying films is a more advanced and micro-level approach because numeric and categorical factors alone do not reveal the different dimensions of films while textual data does. As discovered in the previous EDA section, many of Disney Plus's films are categorized in several genres, indicating a limitation of current genres. Much more detailed and descriptive information about films is required in order to have an in-depth understanding of the films. This is a fundamental knowledge to run any business. Moreover, results of our analysis can be utilized in many areas of business such as product management, product development, marketing, and brand analysis, which all are essential aspects of strategic management. For example, Disney plus can also establish its product development strategy with this text-mined data. By looking at the clusters created by topic modeling, we can see which clusters have the most views, highest ratings, etc. This information will allow Disney Plus to further expand its products that belong to the best performing topics.

Furthermore, using hierarchical clustering, we have established a text-based measure of films that consumers of OTT platforms have much interest in: diversity of films. While current measures such as genre attempt to classify and explain the content of films, oftentimes it is confined to superficial explanations. On the other hand, the dissimilarity measure based on heights score of hierarchical clustering takes into account the latent and dimensional information of the film's content. In addition, establishing the dissimilarity measure allows for a comparison across different OTT platforms, which was not possible before. In this paper, we compared the dissimilarity measures of Disney Plus and Hulu and identified that Hulu has more variety in its film offerings. Using the dissimilarity measure, Disney plus can set a benchmark for its film variety based on other competing platform's dissimilarity scores.

APENDIX

Cluster centers of Disney Plus

```
> Smultiple_kmeans$centers
        type       rated    runtime imdb_rating  imdb_votes months_from_latest        wins language_E
1 -0.4538037 -0.32119131  0.7701751 -0.42393094 -0.08098846         -0.2678485 -0.17599380  0.9770774
2  1.7967070  0.78876892 -0.3685476  0.17293079 -0.10767533         -0.4880093 -0.10918485  0.9432624
3 -0.5302052 -0.09458821 -1.4542367  0.02532076  0.21620755         -0.2986941  0.06424468  0.9482759
4 -0.4492892 -0.54589368  0.4696585  0.53634282 -0.05706849          2.0330629 -0.21675050  0.9848485
5 -0.3151999 -0.34159792 -0.8676631  1.40247482  0.48525978         -0.4808620  5.25687719  1.0000000
  language_M language_S  language_J  language_F language_AR language_CH language_KR language_AL language_BSL
1 0.01719198 0.03151862 0.002865330 0.034383954 0.005730659 0.008595989 0.005730659  0.00286533   0.00000000
2 0.00000000 0.02836879 0.014184397 0.028368794 0.007092199 0.007092199 0.007092199  0.00000000   0.00000000
3 0.00000000 0.05172414 0.011494253 0.074712644 0.011494253 0.000000000 0.005747126  0.00000000   0.01724138
4 0.00000000 0.01515152 0.007575758 0.007575758 0.000000000 0.000000000 0.000000000  0.00000000   0.00000000
5 0.09090909 0.18181818 0.090909091 0.181818182 0.045454545 0.000000000 0.045454545  0.00000000   0.00000000
  language_ASL language_CT language_GM  language_T  language_H  language_D language_CZ language_IT
1  0.000000000  0.00286533 0.011461318 0.000000000 0.000000000 0.000000000 0.000000000 0.011461318
2  0.000000000  0.00000000 0.014184397 0.000000000 0.007092199 0.000000000 0.007092199 0.007092199
3  0.005747126  0.00000000 0.051724138 0.005747126 0.017241379 0.005747126 0.000000000 0.022988506
4  0.007575758  0.00000000 0.007575758 0.000000000 0.000000000 0.007575758 0.000000000 0.007575758
5  0.000000000  0.00000000 0.090909091 0.090909091 0.090909091 0.000000000 0.045454545 0.045454545
  language_RS language_RO language_X language_GR language_HW language_IN language_BG language_AK language_KG
1 0.002865330 0.000000000 0.00000000 0.002865330 0.005730659 0.002865330  0.00000000 0.000000000  0.00000000
2 0.007092199 0.000000000 0.00000000 0.000000000 0.000000000 0.007092199  0.00000000 0.007092199  0.00000000
3 0.011494253 0.005747126 0.01149425 0.005747126 0.000000000 0.000000000  0.00000000 0.000000000  0.00000000
4 0.007575758 0.000000000 0.00000000 0.000000000 0.000000000 0.000000000  0.00000000 0.000000000  0.00000000
5 0.090909091 0.000000000 0.13636364 0.045454545 0.000000000 0.000000000  0.04545455 0.000000000  0.04545455
  language_HB language_SW language_SH language_PO language_PT language_TH language_IK language_SB language_TL
1  0.00000000 0.002865330 0.000000000 0.000000000 0.005730659 0.000000000  0.00286533  0.00286533 0.000000000
2  0.00000000 0.000000000 0.000000000 0.000000000 0.014184397 0.007092199  0.00000000  0.00000000 0.007092199
3  0.00000000 0.005747126 0.005747126 0.005747126 0.000000000 0.000000000  0.00000000  0.00000000 0.000000000
4  0.00000000 0.007575758 0.000000000 0.000000000 0.022727273 0.000000000  0.00000000  0.00000000 0.000000000
5  0.04545455 0.045454545 0.136363636 0.000000000 0.000000000 0.000000000  0.00000000  0.00000000 0.000000000
  language_IR language_YD language_LT language_NW language_TB language_MG language_KZ language_NM language_HG
1 0.000000000 0.000000000  0.00286533 0.005730659 0.000000000 0.000000000 0.000000000  0.00000000 0.000000000
2 0.000000000 0.000000000  0.00000000 0.007092199 0.007092199 0.007092199 0.007092199  0.00000000 0.000000000
3 0.000000000 0.005747126  0.00000000 0.005747126 0.000000000 0.000000000 0.000000000  0.00000000 0.005747126
4 0.007575758 0.000000000  0.00000000 0.000000000 0.000000000 0.000000000 0.000000000  0.00000000 0.000000000
5 0.000000000 0.000000000  0.00000000 0.045454545 0.000000000 0.000000000 0.000000000  0.04545455 0.000000000
  language_ZL language_UK language_PR language_VT language_UD         1         2         3         4
1 0.000000000  0.00286533 0.000000000 0.000000000 0.000000000 0.2854924 0.20408990 0.1339224 0.2274518
2 0.000000000  0.00000000 0.000000000 0.000000000 0.000000000 0.1825647 0.19451147 0.3505124 0.1324846
3 0.005747126  0.00000000 0.005747126 0.005747126 0.005747126 0.1563870 0.27560455 0.1867187 0.2290915
4 0.000000000  0.00000000 0.000000000 0.000000000 0.000000000 0.2508352 0.18977965 0.2959603 0.1252115
5 0.045454545  0.00000000 0.000000000 0.000000000 0.000000000 0.2091183 0.03659787 0.3425048 0.3131383
```

REFERENCES

Alaham, Sanjana. *HULU Shows*. 2021. *Kaggle*, https://www.kaggle.com/sanjanaalaham/hulu-shows/version/1?select=HuluRaw.csv. Accessed 2021.

Blattmann, Josefina. “Netflix: Binging on the Algorithm | by Josefina Blattmann.” *UX Planet*, 2 August 2018, https://uxplanet.org/netflix-binging-on-the-algorithm-a3a74a6c1f59. Accessed 22 November 2021.

Choi, Jaeho, et al. “Using machine learning to revisit the diversification–performance relationship.” *Wiley*, 2021.

Dabbura, Imad. “K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.” *towards data science*, 17 9 2018, https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. Accessed 21 11 2021.

Fontes, Raphael. 2020. *Kaggle*, Raphael Fontes, https://www.kaggle.com/unanimad/disney-plus-shows.

Mahendru, Khyati. “How to Determine the Optimal K for K-Means?” 17 6 2019, https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb. Accessed 21 11 2021.

Moore, Kasey. “The Netflix ID Bible – Every Category on Netflix in 2021.” *What's on Netflix*, 29 September 2021, https://www.whats-on-netflix.com/news/the-netflix-id-bible-every-category-on-netflix/. Accessed 22 November 2021.

Statista Research Department. “Number of OTT video users worldwide from 2017 to 2025.” *statista*, 5 7 2021, https://www.statista.com/forecasts/1207843/ott-video-users-worldwide. Accessed 21 11 2021.

Stoll, Julia. “Share of subscription video-on-demand (SVOD) subscribers who also subscribe to other services in the United States as of December 2020, by service.” *statista*, Julia Stoll, 5 11 2021, https://www.statista.com/statistics/778912/video-streaming-service-multiple-subscriptions/. Accessed 21 11 2021.

Tan, Chade-Meng, et al. “The use of bigrams to enhance text categorization.” *Information Processing & Management*, 2002, p. 529. *ScienceDirect*, https://www.sciencedirect.com/science/article/pii/S0306457301000450.