ISSN:2630-6212

SOCIAL SCIENCE DEVELOPMENT JOURNAL

SSDjournal

Open Access Refereed E-Journal & Refereed & Indexed

http://www.ssdjournal.org / journalssd@gmail.com

Article Arrival Date: 18.06.2023 Published Date: 15.07.2023

Doi Number : http://dx.doi.org/10.31567/ssd.931

Vol 8 / Issue 38 / pp: 187-202

TAVSİYE SİSTEMLERİ ALGORITMALARI KULLANILARAK BIR DİJİTAL PLATFORMUN İÇERİKLERI İÇİN YENİ ÖNERİLER GELİŞTİRİLMESİ

DEVELOPING NEW SUGGESTIONS FOR THE CONTENTS OF A DIGITAL PLATFORM USING RECOMMENDATION SYSTEMS ALGORITHMS

Asst. Prof. Dr. Şeyma BOZKURT UZAN

İstanbul Gelişim Üniversitesi, İktisadi, İdari ve Sosyal Bilimler Fakültesi, suzan@gelisim.edu.tr

İstanbul / Türkiye

ORCID: 0000-0003-3527-3730

Kutluk ATALAY

İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, kutluk.atalay@ogr.iu.edu.tr

İstanbul / Türkiye

ORCID: 0000-0001-7441-5696

ÖZET

Son zamanlarda teknoloji ve bilgi sürekli olarak kendini yenilemektedir. Bu nedenle tüm sektörler doğru bilgiye ulaşmak için teknoloji sayesinde sayısız veriye ulaşabilmektedir. Bu kadar çok veri yığını arasından doğru, anlaşılır ve analiz edilebilir veriler seçebilmek oldukça zor bir konudur. Bu çalışmada da Kaggle adlı veri bilimi platformu tarafından paylaşılan veriler kullanılarak büyük veri analizi ve tahminleme süreçleri analizi gerçekleştirilmiştir. Ortaya çıkan analiz sonuçları sayesinde Netflix'de yer alacak yeni yapımların çalışmada yer alan parametreler bazında değerlendirilmesi mümkün olabilmektedir. Bunun yanında çalışmanın geliştirilmesine yönelik olarak tavsiye sistemleri için farklı ölçütler ve yöntemlerin kullanılabilir olduğu literatürde de görülmüştür.

Anahtar Kelimeler: Büyük Veri Analizi, Kaggle, Tavsiye Sistemleri, Netflix.

ABSTRACT

In recent years, machine learning applications are being used in almost all areas of lives. The main benefits of using machine learning in marketing can be exemplified as follows; content creation, marketing budget optimization and recommendation systems. Recommendation systems are very important and useful when it comes to retaining the current customer. With the help of recommendation systems, companies can retain their customers by recommending their own products, services and contents. In this study, text mining, forecasting processes were carried out using the Netflix contents dataset shared by the data science platform called Kaggle. TfidVectorizer function was used to deal with text data while creating recommendation systems. Two different recommendation systems functions were created in this study.

While first recommendation system function performs only based on title feature of the Netflix contents dataset, the second recommendation system function performs with title, director, cast, listed_in and description features. Thanks to the results of the analysis, it is possible to evaluate the new productions on Netflix on the basis of the features of Netflix contents dataset included in the study. The proposed recommendation system functions provide greater prediction accuracy than conventional systems in data mining. Espicially the recommendation system function that has been developed secondly with the name “get_recommendation_new” uses all features in Netflix contents dataset to recommend new contents to the users.

Keywords: Content-Based Filtering, Data Mining, Machine Learning, Netflix, Recommendation Systems.

1. Introduction

Recently technology and information are renewing themselves constantly. The information that was used continuously in the past and that brought it to the right path every time may not reach the desired result when it loses its validity today. When it comes to undesired result or an negative situation, the most important point is the ability to make decision quickly with the right information. For this reason, all industries using technology to access countless data in order to reach right information. It is very difficult to get the accurate, understandable analyzable data among so many datas. This is the most important step of data science when it comes to information discovery process. The best result in many options has to be discovered and with using this discovery, analyzes should be carried out by keeping up with the developments in technology.

Today, the process of transferring of datas from digital platforms is constantly increasing. Accessing to data is getting easier day by day. The number of websites that share open source datasets are increasing. Analyzing with this much data, supports the accuracy of the results. In this study, Netflix content dataset was used for analyzing, and creating an recommendation system function. Netflix content dataset was downloaded from Kaggle, a subsidiary of Google LLC, an online community of data scientists and machine learning practitioners.

2. Literature Review

Data

Depending on the technological developments, data production is increasing exponentially every day. All humans generate hundreds of data in seconds. Every touch, every click, every photo sharing, every tweet, etc. Numerous data are produced during the day. Data is defined as raw information that does not make sense or cannot be used on its own, but needs to be associated, grouped, interpreted, interpreted and analysed, which forms the basis of information and knowledge [1].

Thanks to the use of the Internet, the data produced is increasing exponentially. At the same time, depending on people's interactions, social media and internet-connected devices are among the factors that increase data production. In the digital world, where heaps of data are formed, these piles of data are stored or flowed instantly. The places where data stacks are stored are called data centers.

Big Data

Big data is a general description of data recorded in digital media and has an organic and global characteristic. The spectrum of Big Data covers a very wide area such as social platform data, e-mails, patient monitoring data, camera recordings, cookies, messages, search engines. However, not every data that people create in social life is included in the scope of Big Data. The concept of big data basically emerged with the “explosion” of global data and was used to describe these enormous data sets.

Compared to traditional datasets, big data generally includes chunks of unstructured/unprocessed data that need more real-time analysis. However, big data opens up new opportunities to discover new data values, helps in-depth analysis of confidential data values, and poses new challenges in how such data can be used and managed effectively. Despite all these opportunities and challenges, big data attracts great attention in every field, from industry to education, from health to government institutions, from science to art [2].

With the increasing popularity of big data and its analysis concepts, etymological studies have been conducted and some academic and non-academic studies have been put forward on this subject. One of them was created by Steve Lohr in 2013 in the New York Times newspaper. During research on the subject, he met with John Mashey, chief scientist at Silicon Graphics in the 1990s. In this interview, Mashey mentioned that he used the term "big data" many times in his presentations to promote his products in the 1990s. In addition, Mashey said that the term “big data” is a fairly simple term and has no claim to fame. he explained. Today, the concept of “Big Data” has emerged as a result of the sharing of data from personal and corporate uses with the whole world through the internet. Big data is the aggregated data that is analyzed and brought to a meaningful level after classification [3].

The concept of Big Data is explained in the literature with the concept of "quantitative thinking". Quantitative thinking actually reveals the decisiveness of technology. Quantitative meanings are obtained thanks to the endless data in the world. Thanks to these meanings, both quantitative thinking and technology become extremely important. also, over-reliance on technology undermines our ability to reason meaningfully about the world [4].

Technology carries and reflects many characteristics of the society in which it is developed, namely power, economy, identity and prejudices. For this reason, we can not talk about a complete change in the social processes that develop, shape and determine the purpose of technology. These social processes show only gradual change. Therefore, technology also carries the characteristics of the society in which it was developed. Understanding Big Data is not only about learning the formal features, but critically approaching Big Data also includes understanding the relationship between variable technology and social functioning [5].

Until recently, bulk data was limited to databases or spreadsheets. With the development of technology, the volume of the data has started to develop rapidly [6]. This bulk data used for analysis, consists of interactions in social media accounts with millions of users worldwide, transactions with internet banking, search engine usage, blogs, e-mails, data collected from sensors on the street or in electronic devices. Big data, on the other hand is a structure which is fed by this bulk of data and also can be processed and can be gotten meaningful outcomes by users [7]. It does not seem possible to limit the use of big data to a particular area. It is used in every field and industry today. Big data features are widely used by government agencies and private companies. One of the areas in which companies use big data the most is to increase their sales by performing customer profile analysis. In this way, it can determine its strategies according to the results of the analysis [8].

According to researches, institutions and companies that used big data a lot in recent years; made approximately 50% more profit, showed 45% more impact in market and sector studies, and reduced their advertising costs by 40% [9]. The impact of data in the world is increasing day by day. The fact that people come across the datas they produce again thanks to the suggestion systems brings both surprise and fear. Thanks to millions of datas produced every day, numerical information about many subjects around the world can be reached. When this is the case, it becomes important to use these datas correctly and beneficially. If the data is used and interpreted correctly, the analyzes obtained in this way will lead to the right decisions for many sectors [10].

Stakeholders and Characteristics of Big Data

Stakeholders of big data; gatherers, users and producers. Big data collectors decide on which parameters the data will be collected and used. Users, on the other hand, use the data for the purpose. Finally, the producers are the actors who produce data voluntarily or involuntarily. The generated data is available in three forms as structured, semi-structured and unstructured. The data that allows processing on it is called structured data. Data that can be found in any order and structure is called unstructured data. The data obtained over the internet and social media has the characteristics of unstructured data. Semi-structured data, on the other hand, are in a unique order and structure. Data from files such as XML files, RSS data, transaction tables are semi-structured data [11].

Dimensions of Big Data

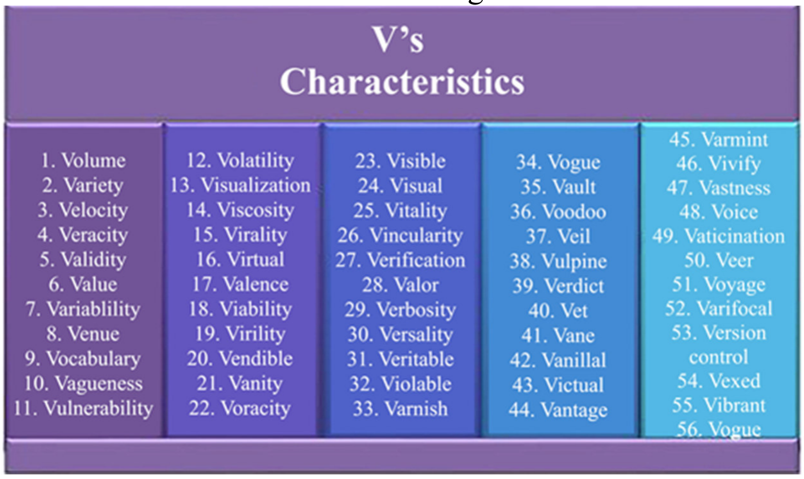
Big data has its own diverse and important features. He identified big data features as volume, variety, and velocity components. The concept of big data has generally been studied by considering these three basic features. These three features are known as the 3V dimension in the literature because English terms start with the letter V [12].

Volume: Even more data is needed to infer more information, and due to such situations, the volume of data increases even more. Therefore, data volume needs to be controlled more effectively. Tiered data storage systems, selective data retention policy, use of statistical sampling, removal of redundant and rarely used data, and outsourcing are important factors in data volume management.

Velocity: Increasing the speed of data generation also requires real-time data control. Accordingly, operational data stores, cached data, point-to-point data routing and decision cycles in data latency are important factors in data rate management.

Variety: In order to achieve more effective analysis results, it requires the use of the same unit values in determining the relationship between data from different fields. In this case, data diversity management aims to collect the data in certain units or to make them suitable for those units.

Table 1. Characteristics of big data: The 56 V's.



V's Characteristics				
1. Volume 2. Variety 3. Velocity 4. Veracity 5. Validity 6. Value 7. Variablility 8. Venue 9. Vocabulary 10. Vagueness 11. Vulnerability	12. Volatility 13. Visualization 14. Viscosity 15. Virality 16. Virtual 17. Valence 18. Viability 19. Virility 20. Vendible 21. Vanity 22. Voracity	23. Visible 24. Visual 25. Vitality 26. Vincularity 27. Verification 28. Valor 29. Verbosity 30. Versality 31. Veritable 32. Violable 33. Varnish	34. Vogue 35. Vault 36. Voodoo 37. Veil 38. Vulpine 39. Verdict 40. Vet 41. Vane 42. Vanillal 43. Victual 44. Vantage	45. Varmint 46. Vivify 47. Vastness 48. Voice 49. Vaticination 50. Veer 51. Voyage 52. Varifocal 53. Version control 54. Vexed 55. Vibrant 56. Vogue

Source: Hussein, AFifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2). Journal of Information Security, 11, 304-328. 2020. doi: 10.4236/jis.2020.114019.

As a result of the literature research, the number of these v's is increasing day by day, as there is 3v of big data and 5v of big data. Among the studies on the properties of big data, Hussein (2020) has the most v. In this study, it is claimed that there are 56 v.

In Table 1, these features are written in the title title. All 56 v are separate big data features. Big data is such a vast concept that more than 56 v will be introduced in the coming days [13].

Data Mining

Data mining is a knowledge discovery process. It is based on a system and consists of algorithms. It is the process of applying models and various algorithms on the data in order to understand the processed or unprocessed data and to draw predictable results from them. [14].

Data mining is a multidisciplinary field of science in computer science that involves the computational process of large datasets to discover influential patterns. The advanced level analysis aims to extract information from huge datas and then convert them into an understandable form for everyone. Also, data mining is considered as an important field in which advanced level techniques are applied to extract patterns in datasets. In the literature, a large number of intelligent methods including machine learning technique, artificial intelligence, database systems, statistics and business intelligence have been designed to analyze data [15].

Data mining has a wide range of uses, as well as big data. Both are sub-branches of data science. Data mining is being used in many industries such as; healthcare, technology, military, commercial. Industries which do not use data mining are doomed to lag behind its competitors as time goes by. In the years of digital transformation, data analysis is also required. In this way, companies will be able to survive [16]. There are three important concepts that should be known at the beginning of the data mining knowledge discovery process. It is important to be able to distinguish the definitions of these concepts [17].

[1] Data is one or more sets of information. They are the facts that transformed into a form that computers can store and process. It emerges with the methods of observation, measurement and experiment.

[2] Information is the processing of data on a particular subject. Information is obtained by processing the data and transforming it into a meaningful form.

[3] Knowledge is the facts obtained by formatting and filtering the data that is transformed into information. Information is used in decision making situations.

Data mining is the stage of extracting and processing the data stored in data sources as a result of the accumulation of big data. This process is provided by statistical methods, database systems, various modeling techniques, and various computer programs used to analyze the data that already exists [18].

Data mining is the process of revealing new data types based on various data kept in data warehouses and using them by processing according to needs. Data science has become very popular and interesting in recent years. The most important reason for this is that the unprocessed (raw) data obtained as a result of the emergence of big data is processed in the fastest way to obtain useful data [19].

Recommendation Systems

Recommendation systems are information systems that are used to recommend suitable products, information or services in order to support the decision-making processes of users, in the environment of online stores and streaming services such as online dating sites and many other industry environments [20]. In terms of the way of their work, recommendation systems can be formed as three different forms such as; content-based filtering recommendation systems, collaborative filtering recommendation systems, and hybrid recommendation systems which is the kind that works by combining two kinds together.

a. Content-based Filtering Recommendation Systems

Content-based filtering recommendation systems consider a fixed set of features for each product type and calculate their similarity to products that the user has been interested in in the past to present similar products.

This method is used in systems with limited product or service area, given that it requires each type of product to perform its characterization itself. [21]. A content-based filtering recommendation system simply created by; a right technique for representing the items (in this study, items are Netflix contents) and users' profiles, a method and a strategy for producing recommendations.
Content-based filtering recommendation systems based on the correlation between the contents in the dataset. Content-based filtering recommendation systems use informations about items in the dataset, as they have been represented as features of the dataset, to calculate the measures of similarities between them. For the measures for similarities, Euclidean and Cosine similarity can be used [22].

b. Collaborative Filtering Recommendation Systems

Collaborative filtering recommendation systems, works as identifying groups of people which have same kind of tastes to those of the users and generate recommendations that they liked. Collaborative filtering recommendation systems, can be grouped into 2 different kinds of approaches such as; memory based algorithms and model based algorithms. Memory-based algorithms generate recommendations to users by using the entire rating matrix. Memory-based algorithm uses an aggregation measure, by considering the ratings from other users, for the same item. Model-based algorithms generates recommendations by creating an model to see the relationships between the items. Model-based algorithms uses several techniques, from unsophisticated techniques such as Naïve Bayes to very sophisticated techniques such as techniques based on aspect models [23].

c. Hybrid Recommendation Systems

Hybrid recommendation systems are used by using content-based recommendation systems and collaborative filtering-based recommendation systems together. Hybrid approaches to form a recommendation system has three different strategies. The first strategy is separately forming content-based filtering recommendation system and collaborative filtering recommendation systems, then combining them. The second strategy is forming content-based capabilities added collaborative filtering recommendation systems. The third approaches is to unify both approaches together into one model [24].

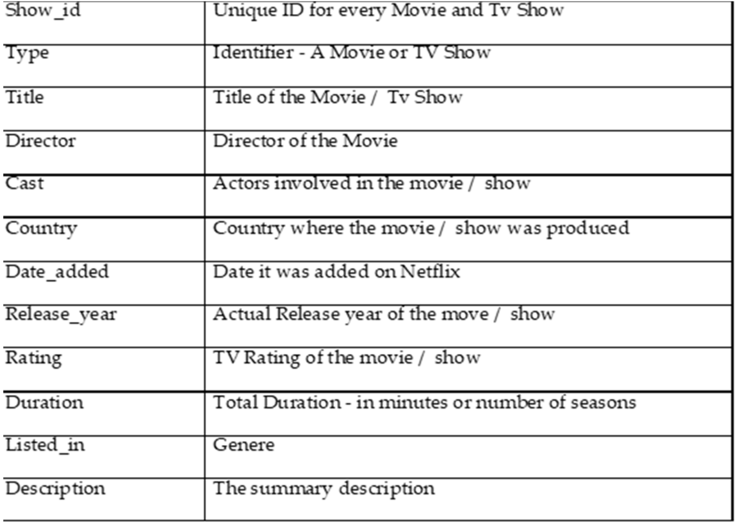
3. Methods

The dataset which was used in this study, accessed via the following link https://www.kaggle.com/datasets/shivamb/netflix-shows . The dataset has 8807 observations and 12 features. In the exploratory data analysis carried out in the data set, it was observed that there were missing datas in the director, cast, country, date_added, rating and duration features. It was decided that the amount of missing data would not adversely affect the performance of the recommendation system function to be created, and data visualization processes were applied. While creating the recommendation system function, content-based recommendation system was chosen as the type of the system. The TfidfVectorizer function was used to work on text data by making use of the scikit-learn machine learning library in the Python programming language.
The productions in the data set were primarily divided into films and shows according to their production types, and exploratory data analysis was applied on the data also with data visualizations. The distribution of the productions according to their content types was analyzed and visualized. In addition, the release years and ratings of the productions were analyzed and visualized in a sequential manner. Then, the missing data in the description feature was filled and the dataset was applied to the Tfidf matrix using the TfidfVectorizer function with the Cosine Similarity criterion.
While Euclidean distance is the most common measure for the distance in recommendation systems function, in this study, another very common method, Cosine similarity was used for the presented recommendation system function.

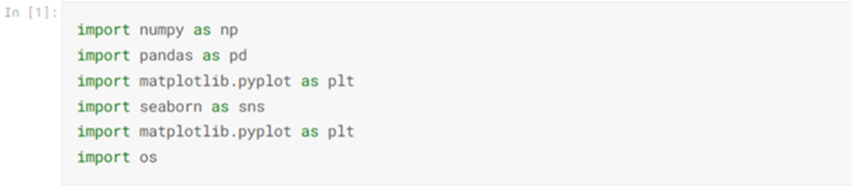
The function named get_recommendation was defined with the argument used as def in the Python programming language, and the recommendation system function was created using the Cosine Similarity criterion. The created function works only based on the description feature in the data set. With the recommendation system function created, recommendations were gotten for the Netflix series Peaky Blinder and Dark.

The function named get_recommendation_new was defined with the argument used as def in the Python programming language, and a new recommendation system function was created using the Cosine Similarity criterion. In this function, title, director, cast, list_in and description features are used together. The created function recalculated the Cosine similarity measure in a different way and used the new Cosine similarity measure to create recommendations. With the new function created, recommendations were created for the Netflix series Dark, Peaky Blinders and Black Mirror.

Table 2. Descriptions of the Features in the Netlflix Content Dataset



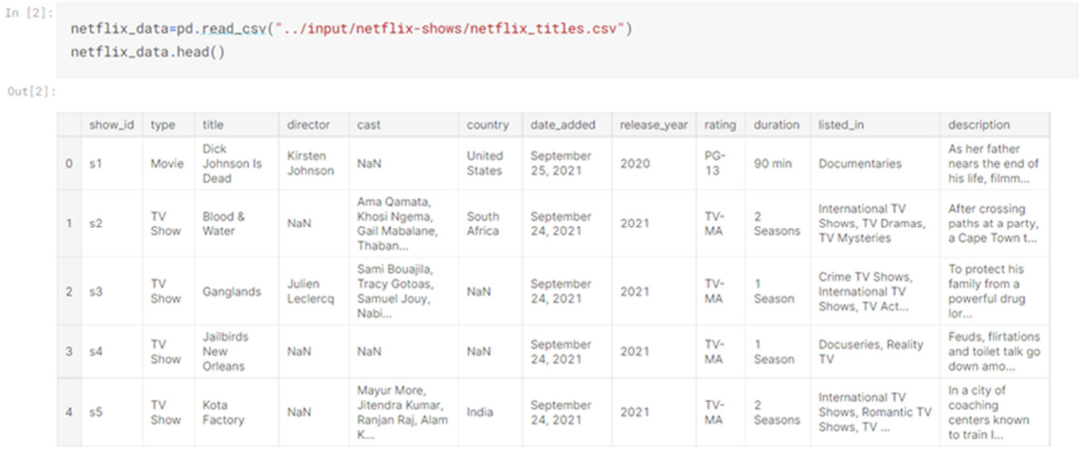
Show_id	Unique ID for every Movie and Tv Show
Type	Identifier - A Movie or TV Show
Title	Title of the Movie / Tv Show
Director	Director of the Movie
Cast	Actors involved in the movie / show
Country	Country where the movie / show was produced
Date_added	Date it was added on Netflix
Release_year	Actual Release year of the move / show
Rating	TV Rating of the movie / show
Duration	Total Duration - in minutes or number of seasons
Listed_in	Genere
Description	The summary description



```
In [1]:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.pyplot as plt
import os
```

Formula 1. Importing Packages into Python Environment

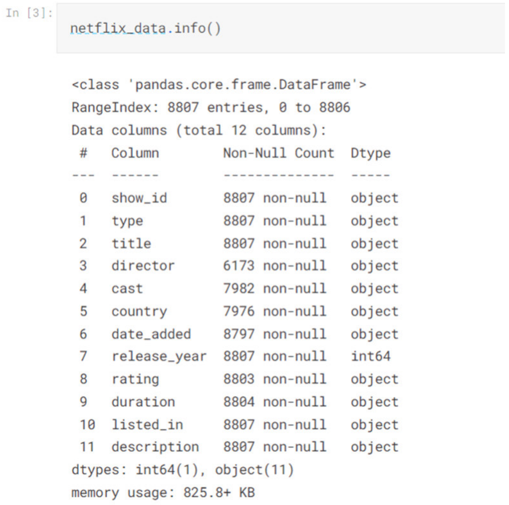
Before starting to analyze Netflix content dataset and creating recommendation system function, first of all, all the packages that required for numerical operations, dataframe operations, visualizations and access to the operator system were imported and activated in the Python notebook. Then, the dataset was imported and by the head() function the first 5 observations were seen in the table below. This step was also used to gain insight about the dataset.



```
In [2]:  netflix_data=pd.read_csv("../input/netflix-shows/netflix_titles.csv")
         netflix_data.head()

Out[2]:
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Formula 2. Exploring The Structure of Netflix Content Dataset

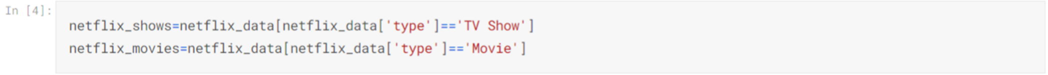


```
In [3]:  netflix_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

Formula 3. Explored Information about the Netflix Content Dataset

When the information about the features of the Netflix content dataset was examined, the inferences below were obtained.

- The Netflix content dataset has 8807 observations and 12 features.
- The feature named "release_year" in Netflix content dataset is in "int64" data type while other features are in object format. Int64 is the short form of integer format, while object format represents strings in Python programming language.
- There are missing values in "director", "cast", "country", "date_added", "rating" and "duration" features of the Netflix content dataset. The "director" feature has the highest number of the missing values.
-

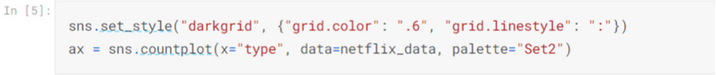


```
In [4]:  netflix_shows=netflix_data[netflix_data['type']=='TV Show']
         netflix_movies=netflix_data[netflix_data['type']=='Movie']
```

Formula 4. Grouping Netflix Content Dataset by Content Types Using the Type Feature

Then, the dataset was divided into two different subgroups according to content types. The first group's content type is "TV Show" and the second group's content type is "Movie".
First group reassigned to a new variable as "netflix_shows" while the second group reassigned to a new variable as "netflix_movies".

Exploratory Data Analysis



```
In [5]:
sns.set_style("darkgrid", {"grid.color": ".6", "grid.linestyle": ":"})
ax = sns.countplot(x="type", data=netflix_data, palette="Set2")
```

Formula 5. Distribution of the Contents of the Dataset by “Type” Feature

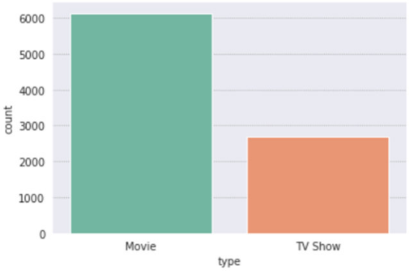
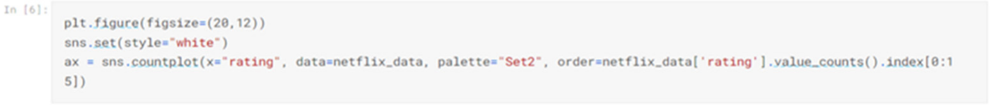


Fig. 1. Visualization of the Distribution of the Contents of the Dataset by “Type” Feature

When the content types in the Netflix content dataset were visualized, it’s seen that the content of the "Movie" type is more than twice as much in the data set as the content of the "TV Show" type.

RATINGS



```
In [6]:
plt.figure(figsize=(20,12))
sns.set(style="white")
ax = sns.countplot(x="rating", data=netflix_data, palette="Set2", order=netflix_data['rating'].value_counts().index[0:1
5])
```

Formula 6. Ordered Ratings of Netflix Contents

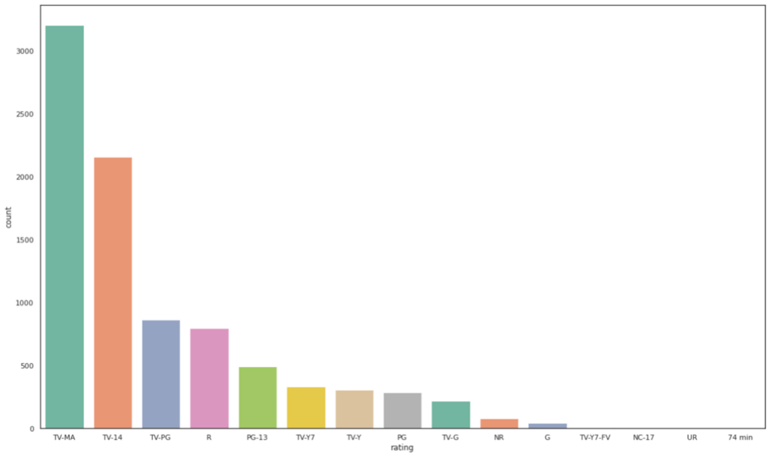
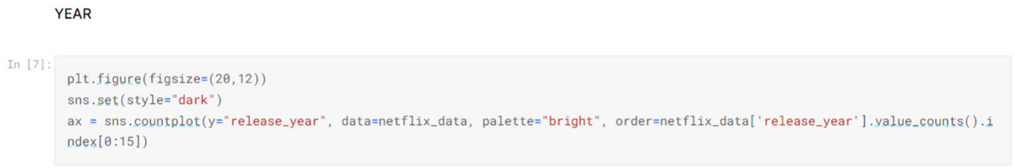


Fig. 2. Visualization of the Ordered Ratings of Netflix Contents

When the contents in Netflix content dataset ordered by the count of content’s rating, it was seen that the “TV-MA” genre received the highest rating. Afterwards, it was seen that the ratings continued to decrease as “TV-14 and TV-PG”.

- TV-MA → Mature Audiences
- TV-14 → Indicates content that is not suitable for viewers aged 14 and under.
- TV-PG → Parental Guideness



YEAR

```
In [7]:
plt.figure(figsize=(20,12))
sns.set(style="dark")
ax = sns.countplot(y="release_year", data=netflix_data, palette="bright", order=netflix_data['release_year'].value_counts().i
ndex[0:15])
```

Formula 8. Distribution of the Contents of the Dataset by “Release_Year” Feature

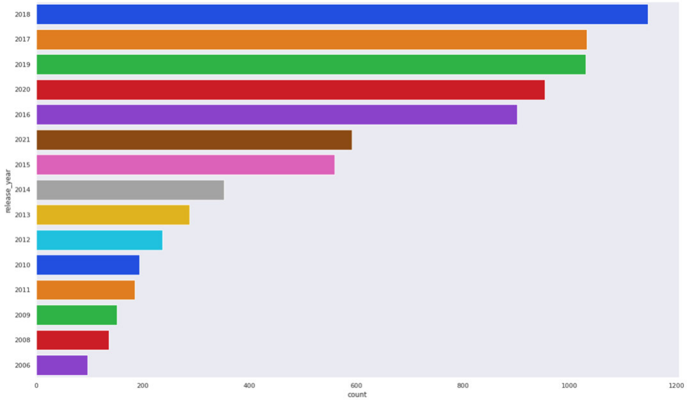
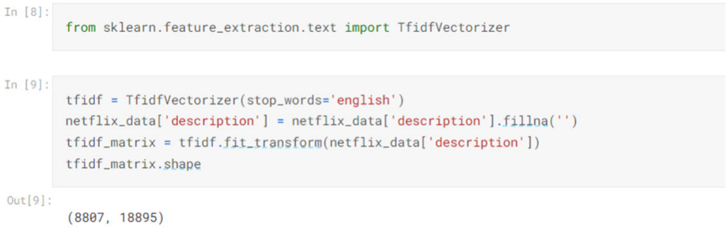


Fig. 3. Visualization of the Distribution of the Contents by “Release_Year” Feature

When the release year of the content on Netflix was examined, it was seen that the most content took place in 2018. It was observed that the ranking continued in 2017 and 2019, decreasingly.
Especially when the number of content released in 2020 and 2021 was examined, it was seen that there was a great decrease. The reason for this can be interpreted as the decrease in many industries due to the Covid-19 pandemic. Because of the Covid-19 pandemic, there are few Netflix contents produced in 2020 and 2021.

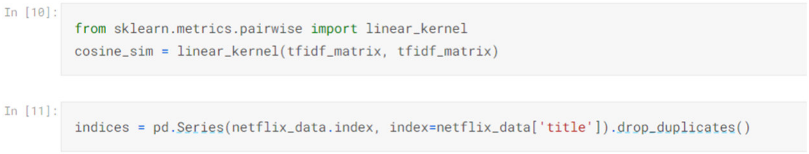
Recommendation System



```
In [8]:
from sklearn.feature_extraction.text import TfidfVectorizer

In [9]:
tfidf = TfidfVectorizer(stop_words='english')
netflix_data['description'] = netflix_data['description'].fillna('')
tfidf_matrix = tfidf.fit_transform(netflix_data['description'])
tfidf_matrix.shape

Out[9]:
(8807, 18895)
```

Formula 8. Importing Libraries and Creating Tf-Idf Matrix for Recommendation System Function

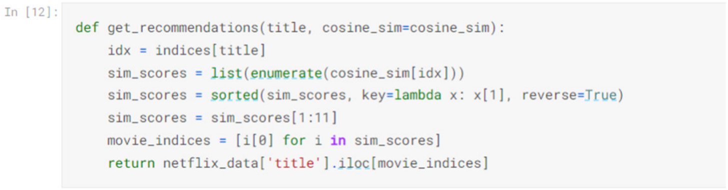
For the recommendation system function, TfidVectorizer function was applied to the text-formatted attributes in the data set, with the stop_words parameter being “English”.
While creating this recommendation system function, Cosine similarity measure was used as similarity measure.



```
In [10]:
from sklearn.metrics.pairwise import linear_kernel
cosine_sim = linear_kernel(tfidf_matrix, tfidf_matrix)

In [11]:
indices = pd.Series(netflix_data.index, index=netflix_data['title']).drop_duplicates()
```

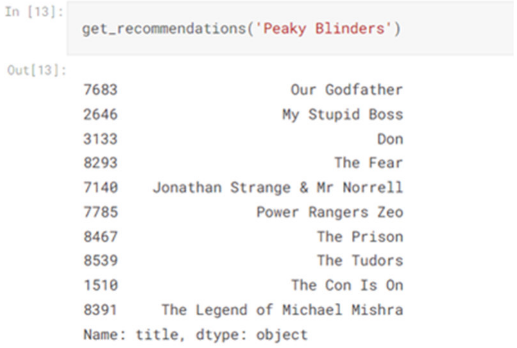
Formula 9. Defining Function for Recommendation System

The cosine similarity measure was applied to the tfidf_matrix transformed above and duplicate observations are dropped from the new dataset. After this stage, the function required for the recommendation system function was defined.



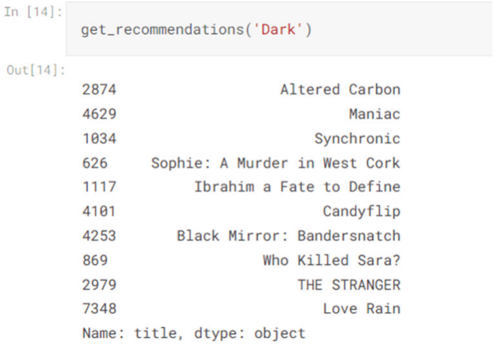
```
In [12]:
def get_recommendations(title, cosine_sim=cosine_sim):
    idx = indices[title]
    sim_scores = list(enumerate(cosine_sim[idx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = sim_scores[1:11]
    movie_indices = [i[0] for i in sim_scores]
    return netflix_data['title'].iloc[movie_indices]
```

Formula 10. Defining Recommendation System Function get_recommendations()



```
In [13]:
get_recommendations('Peaky Blinders')

Out[13]:
7683                        Our Godfather
2646                        My Stupid Boss
3133                                   Don
8293                              The Fear
7140       Jonathan Strange & Mr Norrell
7785                    Power Rangers Zeo
8467                            The Prison
8539                            The Tudors
1510                         The Con Is On
8391          The Legend of Michael Mishra
Name: title, dtype: object
```

Formula 11. Using get_recommendations() function to Get Recommendations for Netflix Series Peaky Blinders in Netflix Content Dataset



```
In [14]:
get_recommendations('Dark')

Out[14]:
2874                   Altered Carbon
4629                           Maniac
1034                      Synchronic
626      Sophie: A Murder in West Cork
1117           Ibrahim a Fate to Define
4101                        Candyflip
4253       Black Mirror: Bandersnatch
869                   Who Killed Sara?
2979                     THE STRANGER
7348                        Love Rain
Name: title, dtype: object
```

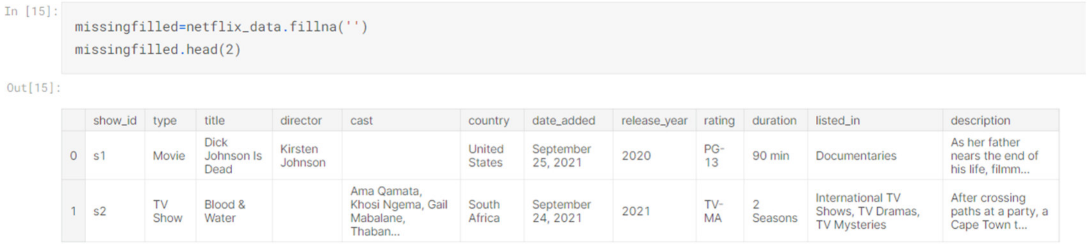
Formula 12. Using get_recommendation() function to Get Recommendations for Netflix Series Dark

By using this function, the content that the audience may like is suggested based on the movies they watch. After defining the function, it was seen that 10 different productions were recommended when the Netflix series called Peaky Blinders was chosen as an example, and when the recommendations were requested for the watching users.

With the same recommendation system function, it was seen that 10 different productions were recommended when asking for recommendations for users watching the Netflix series Dark.

Recommendation system functions may be asked to recommend content based on multiple features. The following procedures were carried out to create a recommendation system function using all the "Title", "Cast", "Director", "Listed in" and "Plot" features in the dataset.

First of all, observations containing NULL values are converted to empty string observations.

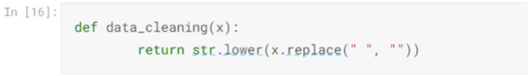


```
In [15]:
missingfilled=netflix_data.fillna('')
missingfilled.head(2)

Out[15]:
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson		United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water		Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...

Formula 13. Creating a Recommendation System Function Using All Features in the Dataset

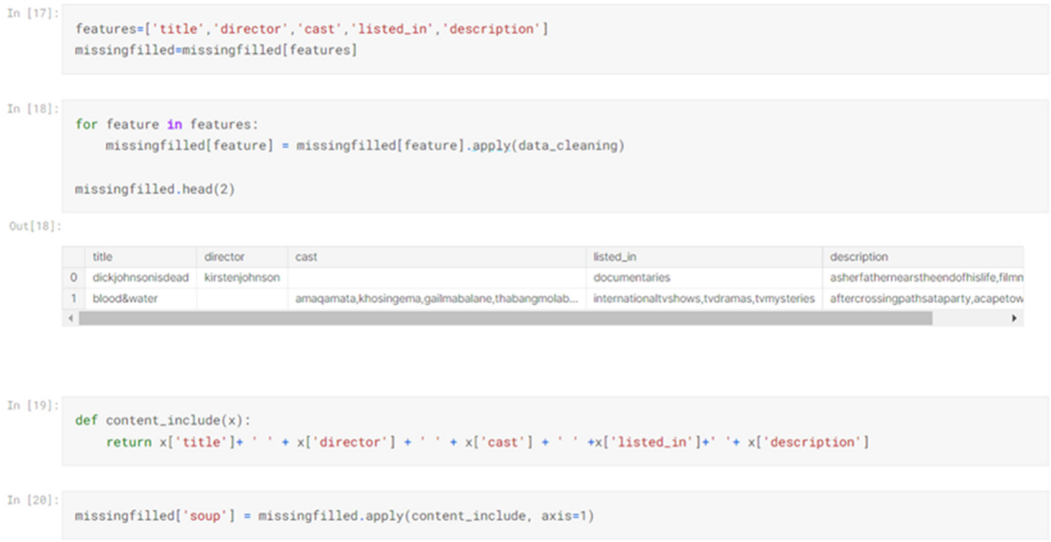
Afterwards, the conversion process was applied so that all the characters in the strings were written in lowercase.



```
In [16]:  def data_cleaning(x):
              return str.lower(x.replace(" ", ""))
```

Formula 14. String Transforming

In the next step, the features of the Netflix content dataset that will filter the model are selected.



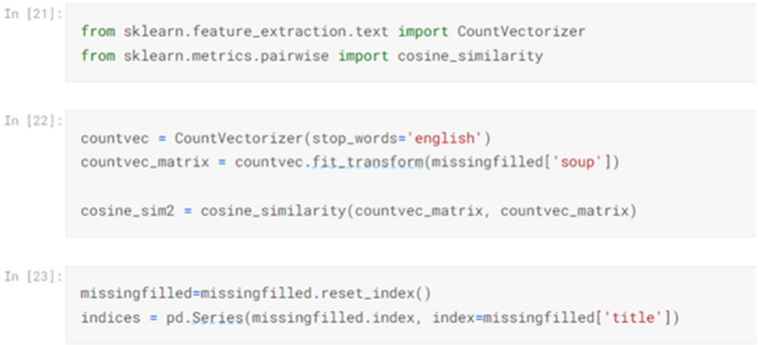
```
In [17]:  features=['title','director','cast','listed_in','description']
          missingfilled=missingfilled[features]

In [18]:  for feature in features:
              missingfilled[feature] = missingfilled[feature].apply(data_cleaning)

          missingfilled.head(2)

Out[18]:
```

	title	director	cast	listed_in	description
0	dickjohnsonisdead	kirstenjohnson		documentaries	asherfathernearstheendofhislife,filmn
1	blood&water		amaqamata,khosingema,gailmabalane,thabangmolab...	internationaltvshows,tvdramas,tvmysteries	aftercrossingpathsataparty,acapetow

```
In [19]:  def content_include(x):
              return x['title']+ ' ' + x['director'] + ' ' + x['cast'] + ' ' +x['listed_in']+' '+ x['description']

In [20]:  missingfilled['soup'] = missingfilled.apply(content_include, axis=1)
```

Formula 15. Feature Selection for Filtering

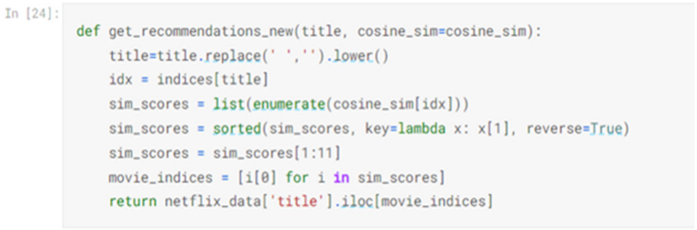
The only thing to do after this stage is to create a new recommendation system function with all features.



```
In [21]:  from sklearn.feature_extraction.text import CountVectorizer
          from sklearn.metrics.pairwise import cosine_similarity

In [22]:  countvec = CountVectorizer(stop_words='english')
          countvec_matrix = countvec.fit_transform(missingfilled['soup'])

          cosine_sim2 = cosine_similarity(countvec_matrix, countvec_matrix)

In [23]:  missingfilled=missingfilled.reset_index()
          indices = pd.Series(missingfilled.index, index=missingfilled['title'])
```

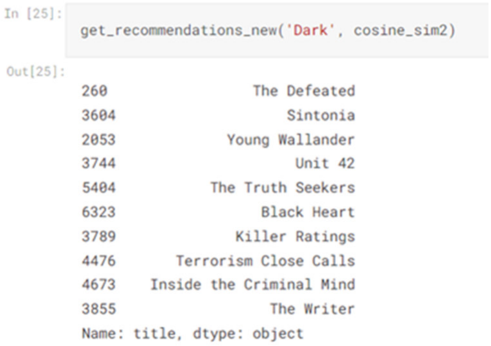
Formula 16. Defining New Recommendation System Function

After selecting the stop_words parameter in the Vectorizer function as "English", the new recommendation system function was created. The new recommendation system function, which has named as get_recommendations_new() is shown below.



```
In [24]:
def get_recommendations_new(title, cosine_sim=cosine_sim):
    title=title.replace(' ','').lower()
    idx = indices[title]
    sim_scores = list(enumerate(cosine_sim[idx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = sim_scores[1:11]
    movie_indices = [i[0] for i in sim_scores]
    return netflix_data['title'].iloc[movie_indices]
```

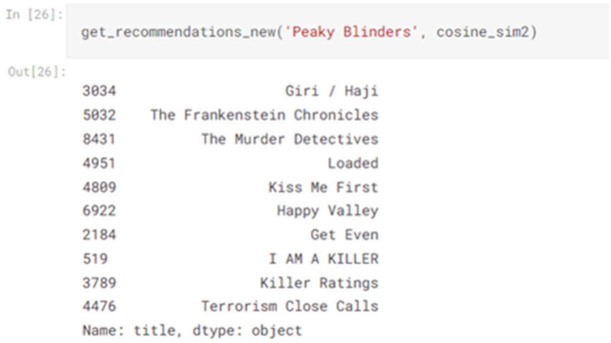
Formula 17. New Recommendation System Function named as get_recommendations_new()



```
In [25]:
get_recommendations_new('Dark', cosine_sim2)

Out[25]:
260                   The Defeated
3604                      Sintonia
2053                Young Wallander
3744                       Unit 42
5404               The Truth Seekers
6323                   Black Heart
3789                Killer Ratings
4476           Terrorism Close Calls
4673         Inside the Criminal Mind
3855                    The Writer
Name: title, dtype: object
```

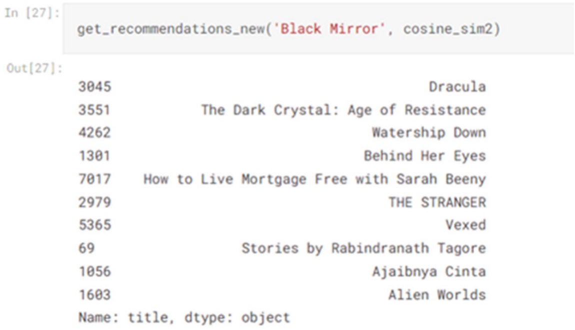
Formula 18. Using get_recommendation new() function to Get Recommendations for Netflix Series Dark

By using new defined recommendation system function, it was seen that 10 different productions were recommended when asked for recommendations for users watching the Netflix series called "Dark".



```
In [26]:
get_recommendations_new('Peaky Blinders', cosine_sim2)

Out[26]:
3034                    Giri / Haji
5032      The Frankenstein Chronicles
8431           The Murder Detectives
4951                         Loaded
4809                    Kiss Me First
6922                   Happy Valley
2184                       Get Even
519                    I AM A KILLER
3789                 Killer Ratings
4476           Terrorism Close Calls
Name: title, dtype: object
```

Formula 19. Using get_recommendation_new() function to Get Recommendations for Netflix Series Peaky Blinders



```
In [27]:
get_recommendations_new('Black Mirror', cosine_sim2)

Out[27]:
3045                                      Dracula
3551           The Dark Crystal: Age of Resistance
4262                               Watership Down
1301                              Behind Her Eyes
7017    How to Live Mortgage Free with Sarah Beeny
2979                                 THE STRANGER
5365                                        Vexed
69                  Stories by Rabindranath Tagore
1056                               Ajaibnya Cinta
1603                                 Alien Worlds
Name: title, dtype: object
```

Formula 20. Using get_recommendation new() function to Get Recommendations for Netflix Series Black Mirror

After the new recommendation system function was defined as get_recommendation_new(), it was used and seen that 10 different productions were recommended when asked for recommendations for users watching the Netflix series "Peaky Blinders".

In addition, the new recommendation system function was defined as get_recommendation_new(), it was used and seen that 10 different productions were recommended when asked for recommendations for users watching the Netflix series "Black Mirror".

Conclusion

In conclusion, this study utilized a dataset comprising Netflix productions to develop two recommendation system functions aimed at suggesting similar productions to Netflix users based on their viewing history. The Cosine similarity measure was employed as the metric for the recommender system functions. The results demonstrated a high level of similarity between the recommendations generated by the developed functions and the productions presumed to be of interest to the audience. Netflix's own recommendation system incorporates three main criteria and three auxiliary criteria, totaling six factors. The main criteria encompass user ratings, shared preferences among users, and genre-based evaluations.

The auxiliary criteria include viewing time periods, devices used for streaming (TV, computer, tablet, mobile phone), and viewing duration. In contrast, this study focused solely on the third main factor employed by Netflix.

One notable distinction between the recommendation system function employed in this study and Netflix's own system is that the former is entirely content-based. The analysis considered various attributes of the productions, including their titles, genres, categories, actors, release years, and ratings. By leveraging this approach, new productions slated for Netflix can be evaluated based on the parameters investigated in this study. Furthermore, the literature suggests that alternative criteria and methods can be explored to enhance recommendation system functions. Moving forward, this study can be expanded by incorporating collaborative filtering methods to form recommendation systems. As the dataset grows, more robust and tailored recommendations can be provided through the development of advanced recommendation system functions. Future research could also explore additional features such as directors or production dates to further enhance the recommendation system. Techniques like the Apriori Algorithm and Eclat Algorithm for Association Rules Mining can be employed to establish relationships between different productions.

Overall, this study sheds light on the effectiveness of content-based recommendation systems using the Netflix dataset and highlights opportunities for further advancements in the field of recommender systems. The findings contribute to the growing body of knowledge in the domain of recommendation systems and serve as a foundation for future research and developments in the field.

Acknowledgements

First and foremost, the authors express their deepest appreciation to Netflix for providing access to the dataset utilized in this study. The collaborative support and cooperation extended by Netflix have been invaluable in facilitating the analysis and generating meaningful insights. Gratitude is also extended to the research participants who graciously shared their viewing history and preferences. Without their voluntary participation, the empirical examination undertaken in this study would not have been feasible. The authors are indebted to their advisors and mentors for their unwavering guidance, expertise, and invaluable contributions throughout the research process. Their profound insights, constructive feedback, and scholarly guidance have been instrumental in refining the study design and methodology.

Furthermore, the authors acknowledge the academic community and researchers in the field of recommendation systems. The extensive body of literature and scholarly works pertaining to this domain has served as a cornerstone for conceptualizing and contextualizing the research framework. The authors are grateful for the wealth of knowledge and insights derived from these scholarly contributions, which have significantly influenced the formulation and execution of this study.

Finally, the authors express their gratitude to their families, friends, and colleagues for their enduring support and encouragement during the course of this research. The authors acknowledge the unwavering motivation, understanding, and assistance provided by their loved ones, which has played a pivotal role in overcoming challenges and maintaining focus on the research objectives. While it may not be possible to individually name all those who have contributed, the authors wish to convey their heartfelt appreciation to each and every individual involved. Their collective efforts and support have been pivotal in transforming this research endeavor into a tangible reality.

References

[1] Yılmaz, Malik. “Enformasyon ve Bilgi Kavramları Bağlamında Enformasyon Yönetimi ve Bilgi Yönetimi.” Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi 49.1 (2009): 95-118.

[2] Chen, Min - MAO, Shiwen - ZHANG, Yin - LEUNG, Victor CM. (2014). Big data: related technologies, challenges and future prospects. Heidelberg: Springer.

[3] Aydın, C., ve Aktaş, B. Talaşlı İmalat Sektöründe Nesnelerin İnterneti ve Anlık Veri Analizi Yöntemleri Kullanarak Üretim Etkinliğinin Artırılması. Istanbul Business Research, 41(1). 2018. https://doi.org/10.26650/ibr.2018.47.01.0002.

[4] Ünal, S. & Sezgin, A. A. Büyük Veri (Big Data)'nin Yapay Zekâ Uygulamalarında Toplumsal Sınıflandırmaya Yönelik Kaygılar . AJIT-e: Academic Journal of Information Technology , 12 (44) , 47-70. 2021. DOI: 10.5824/ajite.2021.01.004.x.

[5] Kitchin, R. Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1). 2014. https://doi.org/10.1177/2053951714528481.

[6] Akkuş, S. Nesnelerin İnterneti Teknolojisinde Güvenli Veri İletişimi—Programlanabilir Fiziksel Platformlar Arasında WEP Algoritması ile Kriptolu Veri Haberleşmesi Uygulaması. Marmara Fen Bilimleri Dergisi, 28(3), 100-111. 2016.

[7] Türk, M. S. Büyük Veri ve Değişim . TRT Akademi , 6 (11) , 5-9 . Retrieved from 2021. https://dergipark.org.tr/tr/pub/trta/issue/60117/870962.

[8] Boz Eravcı, D. Kurumlarin Dijital Dönüşümü: Büyük Veri . Çalışma İlişkileri Dergisi, 11 (1), 90-112. 2020. Retrieved from https://dergipark.org.tr/tr/pub/cider/issue/54745/674025.

[9] Dinç, Y. ve Korkmaz, O. Büyük Verinin Lojistik Sektöründe Kullanimi: Mersin İli Örneği . Verimlilik Dergisi , (4) , 67-88 . 2021. DOI: 10.51551/verimlilik.825813.

[10] Çakırel, Y. İşletmelerde Büyük Veri . Kırklareli Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 5 (1), 52-62. 2016. Retrieved from https://dergipark.org.tr/tr/pub/klujfeas/issue/26888/283283.

[11] Holmes, D. E. Big Data: A Very Short Introduction. 2017. Oxford: Oxford University Press.

[12] Laney, D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group Research Note. 2001. Erişim adresi https://studylib.net/doc/8647594.

[13] Hussein, AFifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2). Journal of Information Security, 11, 304-328. 2020. doi: 10.4236/jis.2020.114019.

[14] Doğan, E. K. ve Şentürk, A. Veri Madenciliği Yöntemleri İle İşveren Sektörünün Sınıflandırılması. Avrupa Bilim ve Teknoloji Dergisi , Ejosat Özel Sayı 2021 (RDCONF) , 227-234. 2021. DOI: 10.31590/ejosat.1039844.

[15] Jayalakshmi, N. ve Mantha Sridevi. 2 - Intelligence methods for data mining task. Artificial Intelligence in Data Mining, Academic Press, 21-39, 2021. DOI : 10.1016/B978-0-12-820601-0.00007-0.

[16] Bayram, S. S. ve Dündar, S. Türkiye'de Banka Şube Lokasyonunun Veri Madenciliği İle Analizi . Uluslararası Bankacılık Ekonomi ve Yönetim Araştırmaları Dergisi , 4 (1) , 34-52 . 2021. DOI: 10.52736/ubeyad.936519.

[17] Kaygısız, E. ve Çağlıyan, V. Bilgi Yönetimi ve Örgütsel Bilgelik İlişkisi Üzerine Sektörel Bir Değerlendirme: Metal ve Makine Sanayi Örneği . Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, (31), 227-240. 2014. Retrieved from https://dergipark.org.tr/tr/pub/susbed/issue/61811/924781.

[18] Akın, B. ve Akın, G. Büyük Veriyle Kaynak ve Kapasite Kısıtları Altında Üretim Planlama ve Çizelgeleme . İzmir İktisat Dergisi , 36 (4) , 759-770. 2021. DOI: 10.24988/ije.718638.

[19] Ulusan Polat, M. ve Babaoğlu, M. Türkiye'de Online Kitap Satışlarının Veri Madenciliği Yöntemiyle Analizi . OPUS International Journal of Society Researches , 17 (36) , 2794-2815 . 2021. DOI: 10.26466/opus.793764.

[20] Son, J., Kim, B. S. Content-based filtering for recommendation systems using multiattribute networks. Expert Systems with Applications, 89, 404-412. 2017. DOI: 10.1016/j.eswa.2017.08.008.

[21] Walek B., Fajmon P. A hybrid recommender system for an online store using a fuzzy expert system. Expert Systems with Applications, 212 (118565), 2023. DOI: 10.1016/j.eswa.2022.118565.

[22] Ghasemi, N. ve Momtazi, S. Neural text similarity of user reviews for improving collaborative filtering recommender systems. Electronic Commerce Research and Applications , (45), 2021. DOI : 10.1016/j.elerap.2020.101019.

[23] Campos, L. M. , Luna, J. F. , Heute, J. , Rueda-Morales, M. Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks. International Journal of Approximate Reasoning, 51, 785-799, 2010. DOI: 10.1016/j.ijar.2010.04.001.

[24] Tian, Y., Zheng, B., Wang, Y., Zhang, Y., Wu, Q. College Library Personalized Recommendation System Based on Hybrid Recommendation Algorithm, Procedia CIRP, 83, 490-494, 2019. DOI : 10.1016/j.procir.2019.04.126.