CSIS 4260 Project Report

Streaming Service Recommendation System

Group: Kunal Ajaykumar Jeshang (300328339) & Blossom Goyal (300347123)

Repository: https://github.com/kjeshang/CSIS4260-StreamingServiceRecommendation

Problem Area

A similarity that many people share is that one of the things they like to do in their spare time is to watch TV shows and movies. In this digital age, a lot of people tend to stream titles online via the internet rather than watch on cable TV. This has given rise to streaming services. These streaming services charge a subscription fee for users to watch titles. Some streaming services are known to produce original titles to have a competitive advantage despite other titles being shared with others. These streaming services are known to have recommendation systems in place that consider the similarity between titles in terms of both content-based & collaborative filtering. However, only titles within a respective streaming service's catalog would be considered; not titles from other streaming services. There are websites freely accessible to everyone that provide a catalog of shows from all streaming services and platforms, but either a robust recommendation system is NOT incorporated into these websites, or not at all. Also, there is a tendency for users to only subscribe to one streaming service when there may actually be other streaming services that have titles they would enjoy watching based on the title/s they like on the streaming service that they are subscribed to.

Therefore, the purpose of this project is to create a recommendation system that draws data of various TV shows & movies from a variety of streaming services that are available in Canada. The system would be able to infer whether they should consider subscribing to other streaming services based on the recommended titles. Despite the project designed in the data analytics perspective, the system would take into consideration a consumer's point of view by providing information on both their selected title and recommended titles.

The project's end result would take the form of a data science application where the user would be able to select the title they enjoy and then receive top ten recommendation results that can be filtered by streaming service as well as adjust results by having a choice in selecting the natural language processing technique. The results would show basic information about the recommedned titles along with the similarity scores. Upon selection of a recommended title from the results, more detailed information can be seen. Also, based on the recommendation results, a simple chart would show the distribution of streaming services. Therefore, the project would have an interactive front-end along with data science processes taking place in the background to output results.

Project Requirements addressed

- Text analysis (understanding the meaning of text but not simple sentiment analysis such as twitter sentiment analysis)
- Analysis of structured / semi structured data (a lot of data set from kaggle will fit into this, but remember this has to be more challenging than 3360 project)
- AI/ML from a variety of data sources (think of combining different data to make a better system)

Dataset

To construct a recommendation system, data of streaming services would be required. Although, before selecting data, identifying streaming services to focus on was paramount. The link below was referred to indentify the most poplular streaming services in Canada that provided solely movies & TV shows. It was identified that Netflix, Amazon Prime, Disney Plus, Paramount, and Apple TV+.
https://techdaily.ca/streaming-services-canada/

Based on the streaming service selected, the corresponding datasets were retrieved from Kaggle. The URLs of the raw datasets are shown below.

- Netflix = https://www.kaggle.com/datasets/shivamb/netflix-shows
- Amazon Prime = https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows
- Disney Plus = https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows
- Paramount = https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows
- Apple TV+ = https://www.kaggle.com/datasets/amritvirsinghx/web-series-ultimate-edition

The raw datasets however were disconnected and not in a structure that would lend well to create a recommendation system. Thus, a significant degree of data cleaning and transformation was conducted to construct a unified dataset that would be proficient for the recommendation system. In terms of data cleaning, certain column values were updated and some rows were removed entirely due to invalid/inconclusive information. In terms of data transformation, unnecessary columns were removed so that only the most informative ones remained that would lend well for identification and text analysis, new columns were made to identify the Streaming Service the titles originate from, and then the raw datasets were combined to make a unified dataset.

After performing the aforementioned, the unified dataset would have to be further transformed to include a column that would be primarily utilized for natural language processing. This new column was named “Textual Info”. The column contained combined values of rating, genre, and description of respective title that were pre-processed for text analysis. The way the row values for the “Textual Info” column are described below.

1. Combine column values rating, genre, and description columns through concatenation.
2. Tokenize the words in the combined column values into a list of elements.

3. Lemmatize the words in the list such that extended words go back to its base form.
4. Update the list such that the elements in the list doesn't include stop words and punctuation
5. Go through each element in the list and remove any occurence of punctuation and special symbols.
6. Combine all of the elements from the list into a line of text.

The data structure of the final unified dataset, which is cleaned, transformed, and pre-processed for text analysis, is reflected in the following table.

Column Name	Detail
type	TV Show or Movie
title	Title of TV Show/Movie
release_year	Year of title's release
rating	Rating of title (e.g., TV-14, TV-MA)
genres	Genre/s of title (e.g., Romantic Movies, Documentaries)
description	Description of the title's plot
Streaming Service	The streaming service where the title can be viewed
Textual Info	Pre-processed textual data for text analysis used the row values of 'rating', 'genres', and 'description' columns

Below is the link of the final unified dataset that is accessible & downloadable from the GitHub repository.
https://github.com/kjeshang/CSIS4260-StreamingServiceRecommendation/blob/main/StreamingServiceRecommendation/streaming_service_titles.xlsx

Below is the link from the GitHub repository that shows the codebase of how the final dataset was prepared.
https://github.com/kjeshang/CSIS4260-StreamingServiceRecommendation/blob/main/StreamingServiceRecommendation/1_preparation.ipynb

Analysis

As explained, the intent of this project is to create a recommendation system. As the descriptions of the titles in the dataset are not extremely detailed, it was decided to implement two text analysis methods on combined information of rating, genre, and description. These methods are explained in the following paragraphs in the context of this project.

“TF-IDF, or term frequency-inverse document frequency, is a figure that expresses the statistical importance of any given word to the document collection as a whole” (Chouinard, 2022). The mathematical expression for TF-IDF can be seen below.

TF-IDF = TF * IDF

- TF (i.e., Term Frequency): Number of times a word appears in a document / number of words in the document.
- IDF (i.e., Inverse Document Frequency): log(Number of documents / Number of documents that contain the word).

Similar to TF-IDF, N-Grams is also used for text mining and natural language processing tasks. It refers to a set of co-occurring words. When using N-Grams you typically move one word forward, but can move any number of words forward (i.e., X words forward). For example, you can move 2 words forward, whereby each pair of words is considered rather than each individual word. In the context of this project, three types of N-grams are considered (Ganesan, 2020).

- Unigrams (single word)
- Bigrams (pair of words)
- Trigrams (tuple of three words)

In the context of the recommendation system, a Document Term Matrix is created via vectorization from the collection of row values in the “Textual Info” column. As the project is a recommendation system, the selected title to be used for comparison is shifted to the top of the dataset prior to vectorization. The values in the matrix are then transformed by being autoscaled via normalization (i.e., mean and standard deviation). Then the matrix is used to construct a pairwise cosine similarity kernel whereby similarity scores between the selected title and the other titles can be calculated. The lower and upper limit of the similarity score will be 0 and 1, respectively. If, a similarity score is above 0.5, it can be considered that the similarity between a user selected title and respective recommended title is relatively high ("What is a good threshold for CosineSimilarity Measure?", 2020). For the purpose of concise-ness, the top ten recommendation results would be output. As two disinct methods are explored in this project, when applying N-Grams the vectorization does not consider the importance of words like in TF-IDF so as to compare the difference of results between the methods.

Below is the link from the GitHub repository that shows how the aforementioned methods work in a simple example using Python.
https://github.com/kjeshang/CSIS4260-StreamingServiceRecommendation/blob/main/StreamingServiceRecommendation/2_analysis.ipynb

Goals & Desired Insights

- Retrieve TV Show and/or Movie recommendations that are not part of the user’s currently subscribed streaming service/s
- Find out the distribution of streaming services amongst the top ten recommendations

- Assess the commonality and differences amongst recommendation results between TF-IDF and N-Grams (Unigrams, Bigrams, Trigrams) methods
- Identify any potential relationships or differences between recommendation results of titles that are of similar/distinct genres (if possible)

User Interface

Rather than utilizing a Jupyter Notebook to display the recommendation output, we thought it would be interesting to create an interactive user interface. This way, one can simply play around with the streaming service recommendation system's various options in terms of textual analysis techniques, streaming service checklist options, and title selections. This way results could be viewed more organically on-the-fly. This lends well in a presentation as an audience would be more likely to be drawn to an interactive user interface rather than a Jupyter Notebook. The user interface was created with the help of the Dash library using the Python programming language. The Dash library is free-to-use and helps to serve as a great front-end for Data Science & Data Analytics oriented applications. Thus, despite conducting a simple study of recommendation systems using some textual analysis techniques, but the output of the study is in the form of a Data Analytics application.

We attempted to deploy this project on Heroku so that is accessible without need of dependency installation. We were successfully able to deploy, but due to using the free tier dynos, the application was not able to perform at the required data processing level to churn out recommendations. This was unfortunate but an interesting experience to attempt this. We understand now that deploying a self-created data science on a free platform would not be optimal. Thus, in a real world environment with the necessary funds and resources, we would deploy this type of application to a cloud platform such as AWS, Google Cloud, or Azure. Using Kubernetes could be helpful here as well.

Data Analytics Model

Out of the four main data analytics models, the project incorporates two of them: predictive analytics and descriptive analytics ("What Are the 4 Main Analytical Models?", n.d.). The components of the project that churn out recommended titles based on a user selected title falls along the lines of predictive analytics. The components of the project that infer potential streaming services for a user to subscribe-to from the recommended titles falls along the lines of descriptive analytics.

Tools & Libraries

The table below describes the tools & technologies used to create this project.

Programming Language	Python 3.8.9+

Integrated Development Environment	Microsoft Visual Studio Code ● Data Preparation, NLP Pre-Processing, and Prototyping: Jupyter Notebook ● Development: Python Script
Operating System	Windows 11
Cleaned Data File Format	Microsoft Excel (.xlsx)
Source of Raw Data Files	Kaggle

The table below describes the Python libraries used to create this project.

Pandas	Used predominatly to interact with data during preparation stage, but also used to perform unsupervised learning for the recommendation system as well as visualization
NLTK	● Corpus module used to retreive stop words that are used for NLP Pre-Processing ● Tokenize module used to convert row values in the "Textual Info" column into a list of elements for NLP Pre-Processing ● Stem module used to lemmatize elements in the above mentioned list (i.e., convert extended words to its base form if necessary) for NLP Pre-Processing
String	Used to retrieve punctuation symbols/characters for NLP Pre-Processing
Re	Used to remove special characters & symbols from individual words for NLP Pre-Processing
Scikit Learn	● Feature Extraction Text ○ TFIDFVectorizer used for TF-IDF based recommendations ○ CountVectorizer used for N-Grams based recommendations ● Metrics Pairwise ○ Linear kernel used for TF-IDF based recommendations ○ Cosine_similarity used for N-Grams based recommendations
Plotly Express	● Used to create visualization of streaming service distribution
Dash	● Dash: used to instantiate front-end of project which takes the shape of a web application ● HTML: used to create some static web elements ● DCC: used to create some interactive web elements and handles structural aspects

	● Input, Output, Callback: used to enact interactivity and output recommendation results ● Dash_table: hold recommendation results in a table
Dash Bootstrap Elements	Used to create interactive-stylistic web elements

Roadblocks

- Slow runtime when processing recommendation results due to large final database. Although, this is dependent on the power of the user's computer.
- The creators of the application are quite inexperienced in textual analysis. Thus, the implementations of TF-IDF and NGrams may not be optimally calibrated to output the maximum-optimal recommendation results. In particular it is their first time tackling a project involving textual analysis and creating a recommendation system. Thus, there could be a variability in the similarity scores.
- As the application uses a collection of datasets for each respective streaming service and combines them together, there may be a disproportionate amount of content from a streaming service that has been established for a long time such as Netflix over that of a newcomer like Apple TV+. A reason for this could be that the established streaming services have more titles, and so less varied distribution of streaming services that could be considered based on the user's selected title.
- The datasets used to create the final unified dataset were created by someone else. Despite, performing data cleaning and transformation, there may be instances where some row values may be invalid or may not be conclusive, but are still considered when outputting recommendation results. This could lead to an invalid title appearing in the top ten recommendation results
- Potential lack of recency of title catalog information in the streaming service datasets, and may be region specific to the person who made the streaming service datasets. For example, the person that made the Netflix raw dataset from Kaggle may have last been updated in 2020, but it is currently 2022. Also the person may be situated in a country; e.g., India. Thus, the title catalog information in the dataset may refer to titles only available in the person's given region; not globally.
- The descriptions of the Movies and TV shows may be too brief, so the recommendations may not make perfect sense or be conclusive.

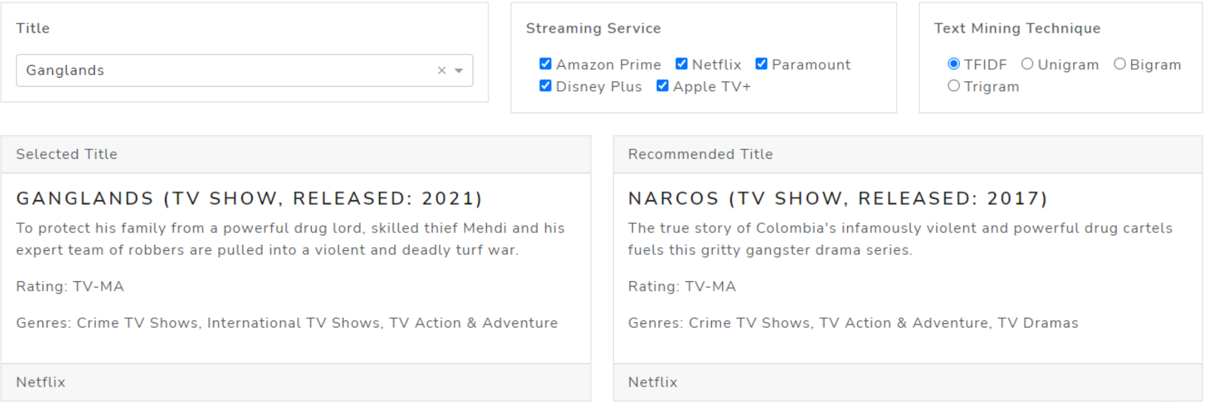
Example Scenarios using the Application

Let us say a consumer in Canada has subscription to only Netflix. It is overall the most popular streaming service in various regions of the world (including Canada) due to its long standing history in its respective industry, thus, it is highly reputable, and also has wide selection of titles in its catalog. They however do not have subscriptions to the other streaming services due to lack of familiarity with their respective title catalog. They would be open to finding out about more titles on other streaming services and consider subscribing to them.

Scenario 1

The consumer is a great fan of the show “Ganglands” on Netflix. They want to find out similar shows regardless if they or the Netflix streaming service or other streaming services that they are not subscribed to. The text mining technique used is TF-IDF; it is the technique selected by default. Below are screenshots of the results.

Selected Title vs Recommended Title with Highest Similarity



Title

Ganglands

Streaming Service

☑ Amazon Prime ☑ Netflix ☑ Paramount
☑ Disney Plus ☑ Apple TV+

Text Mining Technique

● TFIDF ○ Unigram ○ Bigram
○ Trigram

Selected Title

GANGLANDS (TV SHOW, RELEASED: 2021)

To protect his family from a powerful drug lord, skilled thief Mehdi and his expert team of robbers are pulled into a violent and deadly turf war.

Rating: TV-MA

Genres: Crime TV Shows, International TV Shows, TV Action & Adventure

Netflix

Recommended Title

NARCOS (TV SHOW, RELEASED: 2017)

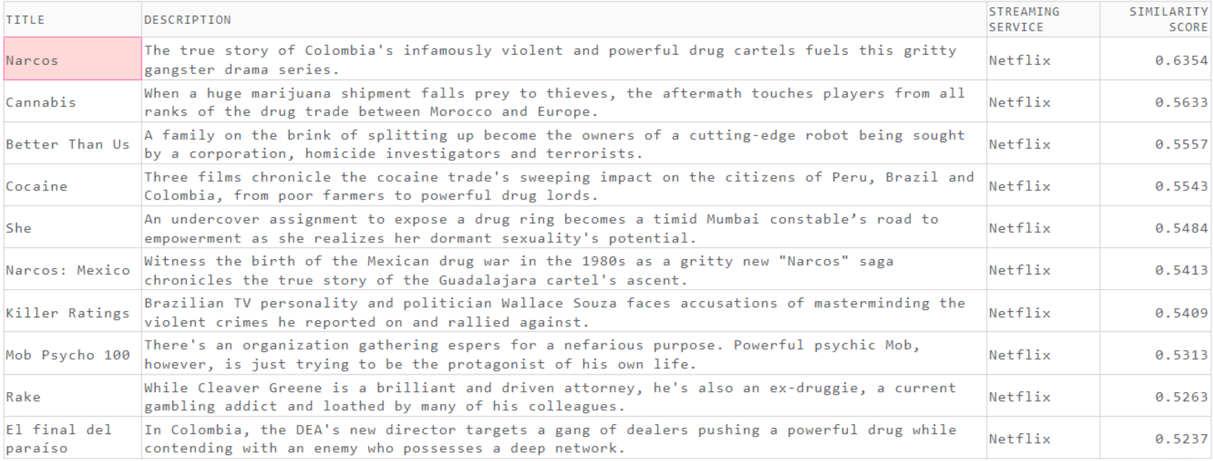
The true story of Colombia's infamously violent and powerful drug cartels fuels this gritty gangster drama series.

Rating: TV-MA

Genres: Crime TV Shows, TV Action & Adventure, TV Dramas

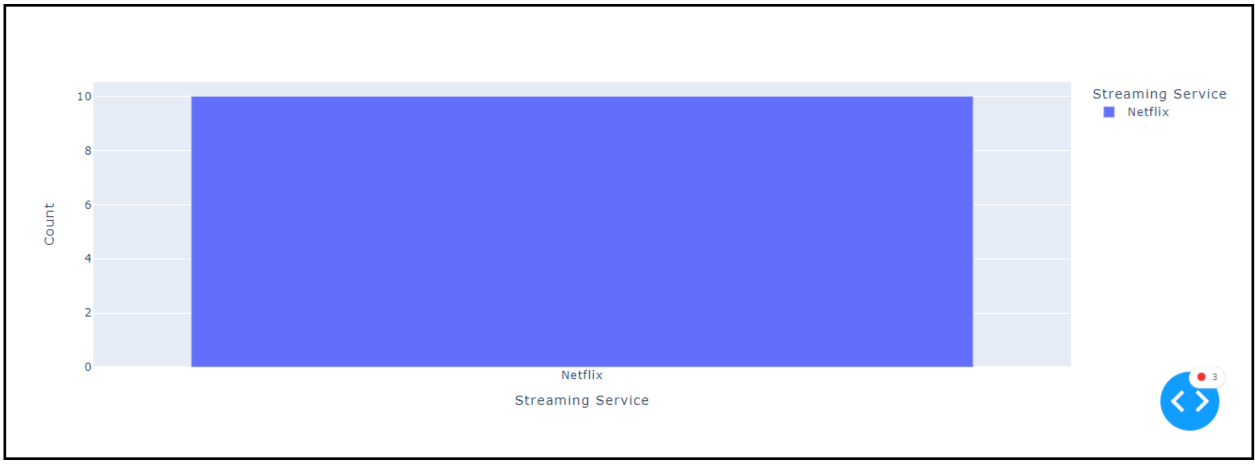
Netflix

Recommendation Results



TITLE	DESCRIPTION	STREAMING SERVICE	SIMILARITY SCORE
Narcos	The true story of Colombia's infamously violent and powerful drug cartels fuels this gritty gangster drama series.	Netflix	0.6354
Cannabis	When a huge marijuana shipment falls prey to thieves, the aftermath touches players from all ranks of the drug trade between Morocco and Europe.	Netflix	0.5633
Better Than Us	A family on the brink of splitting up become the owners of a cutting-edge robot being sought by a corporation, homicide investigators and terrorists.	Netflix	0.5557
Cocaine	Three films chronicle the cocaine trade's sweeping impact on the citizens of Peru, Brazil and Colombia, from poor farmers to powerful drug lords.	Netflix	0.5543
She	An undercover assignment to expose a drug ring becomes a timid Mumbai constable’s road to empowerment as she realizes her dormant sexuality's potential.	Netflix	0.5484
Narcos: Mexico	Witness the birth of the Mexican drug war in the 1980s as a gritty new "Narcos" saga chronicles the true story of the Guadalajara cartel's ascent.	Netflix	0.5413
Killer Ratings	Brazilian TV personality and politician Wallace Souza faces accusations of masterminding the violent crimes he reported on and rallied against.	Netflix	0.5409
Mob Psycho 100	There's an organization gathering espers for a nefarious purpose. Powerful psychic Mob, however, is just trying to be the protagonist of his own life.	Netflix	0.5313
Rake	While Cleaver Greene is a brilliant and driven attorney, he's also an ex-druggie, a current gambling addict and loathed by many of his colleagues.	Netflix	0.5263
El final del paraíso	In Colombia, the DEA's new director targets a gang of dealers pushing a powerful drug while contending with an enemy who possesses a deep network.	Netflix	0.5237

Streaming Service Distribution

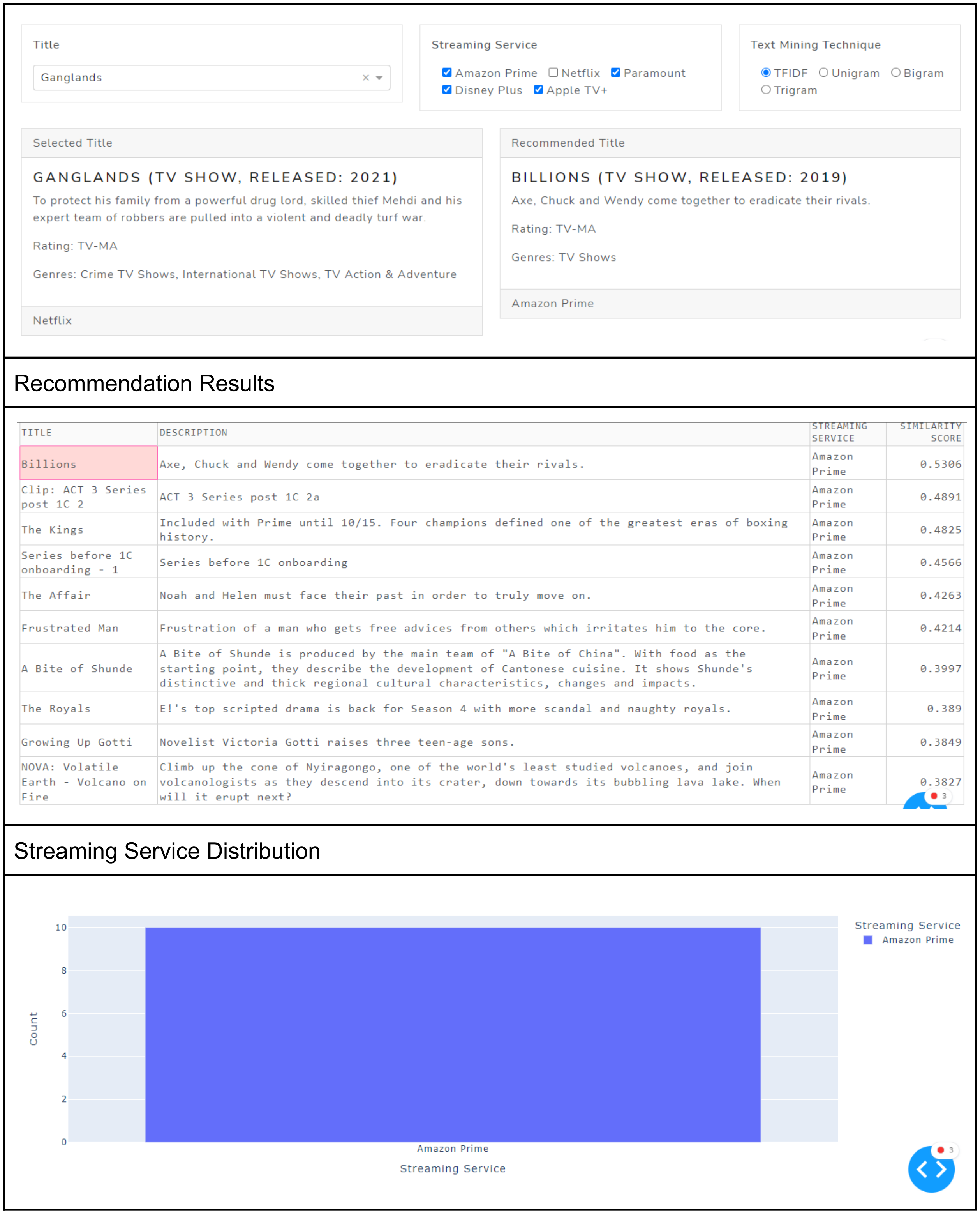


According to the output shown from the screenshots above, it seems that most of the shows that the consumer would enjoy remain on the Netflix streaming service. The "Narcos" TV show has the highest similarity score out of all of the other recommendation results. The fact that "Narcos" is the recommended title with the highest similarity score makes sense as "Ganglands" is also a TV show that is meant for a mature audience. Also, both shows are of the crime, action, and adventure genre, that have violent and gritty content involving gang warefare. For now, the consumer can simply remain on the Neflix streaming service and not spend money on subscribing to another service.

Scenario 2

Continuing on from Scenario 1, the consumer has watched "Narcos" based on the recommendation results discussed above. They have had a lot of free time and ended up watching all of the subsequent recommended titles shown in the screenshots in Scenario 1. This is justified by the fact that all of the similarity scores for the subsequent recommended titles were above 0.5. They want to watch similar shows that exist on other streaming services and will consider subscribing to them. Thus they would use the application again but this time uncheck the 'Netflix' option on the streaming service checklist; the other options remain checked. The same text mining technique would be used as in Scenario 1. Below are screenshots of the results.

Selected Title vs Recommended Title with Highest Similarity



Title

Ganglands × ▾

Streaming Service

☑ Amazon Prime ☐ Netflix ☑ Paramount
☑ Disney Plus ☑ Apple TV+

Text Mining Technique

● TFIDF ○ Unigram ○ Bigram
○ Trigram

Selected Title

GANGLANDS (TV SHOW, RELEASED: 2021)

To protect his family from a powerful drug lord, skilled thief Mehdi and his expert team of robbers are pulled into a violent and deadly turf war.

Rating: TV-MA

Genres: Crime TV Shows, International TV Shows, TV Action & Adventure

Netflix

Recommended Title

BILLIONS (TV SHOW, RELEASED: 2019)

Axe, Chuck and Wendy come together to eradicate their rivals.

Rating: TV-MA

Genres: TV Shows

Amazon Prime

Recommendation Results

TITLE	DESCRIPTION	STREAMING SERVICE	SIMILARITY SCORE
Billions	Axe, Chuck and Wendy come together to eradicate their rivals.	Amazon Prime	0.5306
Clip: ACT 3 Series post 1C 2	ACT 3 Series post 1C 2a	Amazon Prime	0.4891
The Kings	Included with Prime until 10/15. Four champions defined one of the greatest eras of boxing history.	Amazon Prime	0.4825
Series before 1C onboarding - 1	Series before 1C onboarding	Amazon Prime	0.4566
The Affair	Noah and Helen must face their past in order to truly move on.	Amazon Prime	0.4263
Frustrated Man	Frustration of a man who gets free advices from others which irritates him to the core.	Amazon Prime	0.4214
A Bite of Shunde	A Bite of Shunde is produced by the main team of "A Bite of China". With food as the starting point, they describe the development of Cantonese cuisine. It shows Shunde's distinctive and thick regional cultural characteristics, changes and impacts.	Amazon Prime	0.3997
The Royals	E!'s top scripted drama is back for Season 4 with more scandal and naughty royals.	Amazon Prime	0.389
Growing Up Gotti	Novelist Victoria Gotti raises three teen-age sons.	Amazon Prime	0.3849
NOVA: Volatile Earth - Volcano on Fire	Climb up the cone of Nyiragongo, one of the world's least studied volcanoes, and join volcanologists as they descend into its crater, down towards its bubbling lava lake. When will it erupt next?	Amazon Prime	0.3827

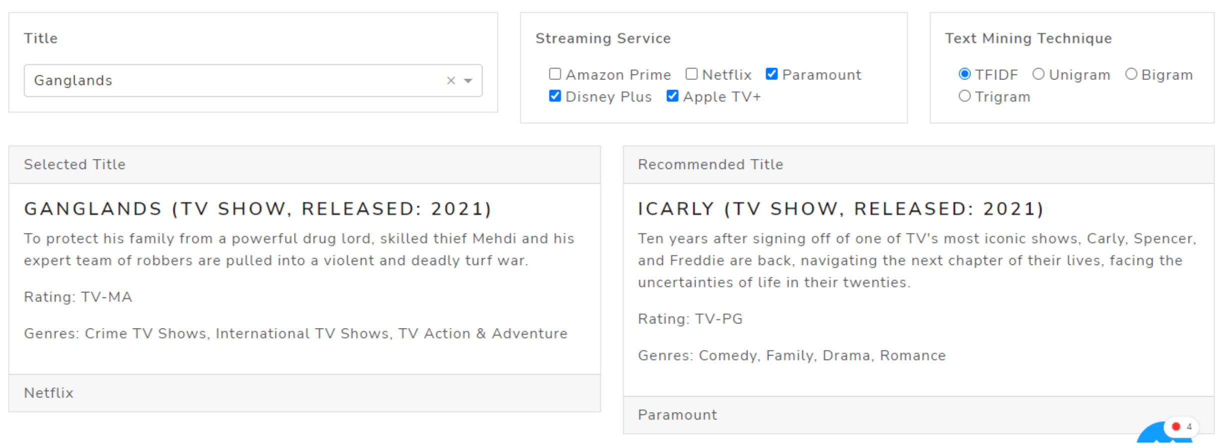
Streaming Service Distribution

According to the output shown in the above screenshots, it seems that most of the shows the consumer would enjoy is on Amazon Prime. The "Billions" TV show is the recommended title with the highest similarity score. It is the only title amongst the recommendation results that has a score above 0.5. Specifically, in “Billions”, a group of individuals want to eradicate their rivals. On “Ganglands”, the show involves a turf war, thus a conflcut, which can correspond to the plot involving the eradication of the protagonists’ rivals in “Billions”. The other recommended titles have a similarity score of below 0.5. The selected title and recommended title are both meant for mature audience. The genres may differ but both shows have violent elements. Thus, it could be suggested that based on the results, the consumer may consider watching “Billions” by subscribing to Amazon Prime’s one month free trial and then cancelling to the subscription altogether after watching the TV show. If they are okay with watching shows that may not have a strong similarity to “Ganglands”, then they can consider subscribing.

Scenario 3

The consumer is feeling experimental from their experience watching the recommended title/s on Amazon Prime. As Netflix and Amazon Prime are well established streaming services in Canada, they want to see what other titles are out there in the newer streaming services such as Paramount, Disney Plus, and Apple TV+. Thus, both Netflix and Amazon Prime options would be unchecked on the Streaming Service checklist; all others would be checked. The same text mining technique would be used as in Scenario 2. Below are screenshots of the results.

Selected Title vs Recommended Title with Highest Similarity



Title

Ganglands

Streaming Service

☐ Amazon Prime ☐ Netflix ☑ Paramount
☑ Disney Plus ☑ Apple TV+

Text Mining Technique

● TFIDF ○ Unigram ○ Bigram
○ Trigram

Selected Title

GANGLANDS (TV SHOW, RELEASED: 2021)

To protect his family from a powerful drug lord, skilled thief Mehdi and his expert team of robbers are pulled into a violent and deadly turf war.

Rating: TV-MA

Genres: Crime TV Shows, International TV Shows, TV Action & Adventure

Netflix

Recommended Title

ICARLY (TV SHOW, RELEASED: 2021)

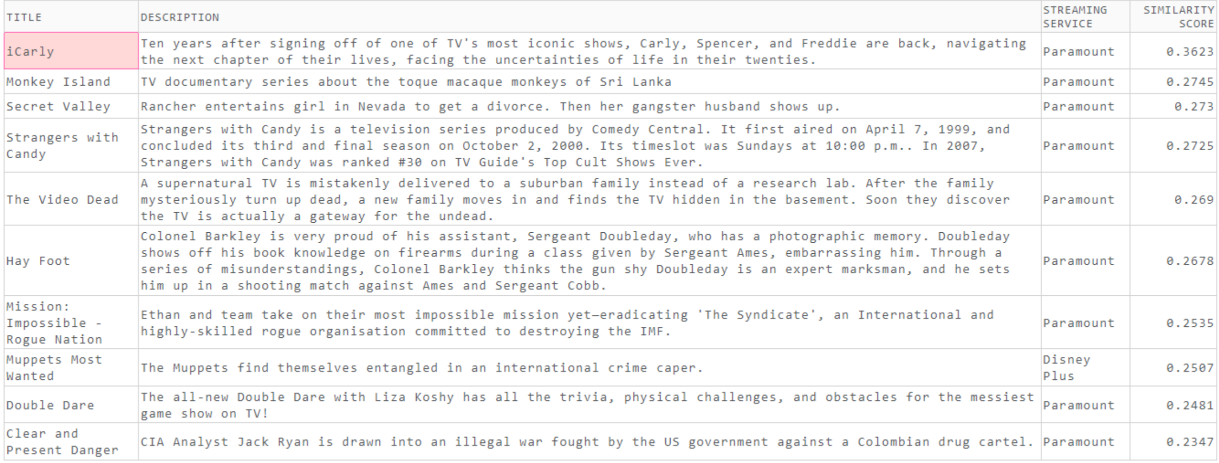
Ten years after signing off of one of TV's most iconic shows, Carly, Spencer, and Freddie are back, navigating the next chapter of their lives, facing the uncertainties of life in their twenties.

Rating: TV-PG

Genres: Comedy, Family, Drama, Romance

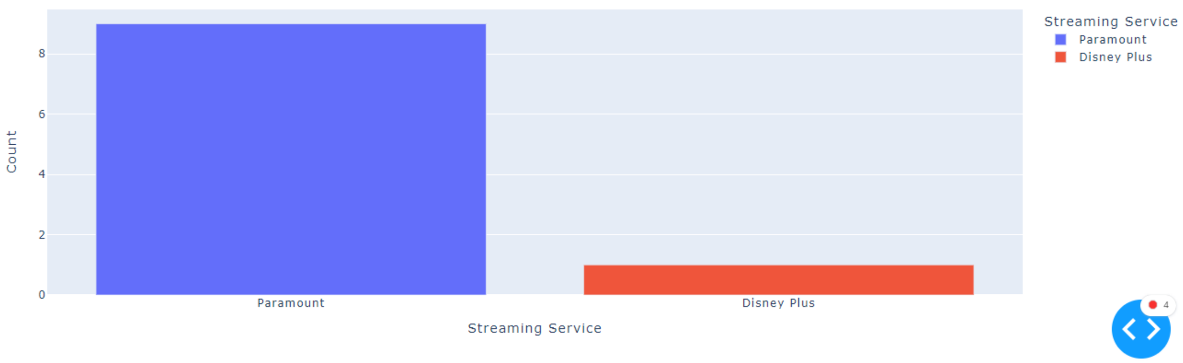
Paramount

Recommendation Results



TITLE	DESCRIPTION	STREAMING SERVICE	SIMILARITY SCORE
iCarly	Ten years after signing off of one of TV's most iconic shows, Carly, Spencer, and Freddie are back, navigating the next chapter of their lives, facing the uncertainties of life in their twenties.	Paramount	0.3623
Monkey Island	TV documentary series about the toque macaque monkeys of Sri Lanka	Paramount	0.2745
Secret Valley	Rancher entertains girl in Nevada to get a divorce. Then her gangster husband shows up.	Paramount	0.273
Strangers with Candy	Strangers with Candy is a television series produced by Comedy Central. It first aired on April 7, 1999, and concluded its third and final season on October 2, 2000. Its timeslot was Sundays at 10:00 p.m.. In 2007, Strangers with Candy was ranked #30 on TV Guide's Top Cult Shows Ever.	Paramount	0.2725
The Video Dead	A supernatural TV is mistakenly delivered to a suburban family instead of a research lab. After the family mysteriously turn up dead, a new family moves in and finds the TV hidden in the basement. Soon they discover the TV is actually a gateway for the undead.	Paramount	0.269
Hay Foot	Colonel Barkley is very proud of his assistant, Sergeant Doubleday, who has a photographic memory. Doubleday shows off his book knowledge on firearms during a class given by Sergeant Ames, embarrassing him. Through a series of misunderstandings, Colonel Barkley thinks the gun shy Doubleday is an expert marksman, and he sets him up in a shooting match against Ames and Sergeant Cobb.	Paramount	0.2678
Mission: Impossible - Rogue Nation	Ethan and team take on their most impossible mission yet—eradicating 'The Syndicate', an International and highly-skilled rogue organisation committed to destroying the IMF.	Paramount	0.2535
Muppets Most Wanted	The Muppets find themselves entangled in an international crime caper.	Disney Plus	0.2507
Double Dare	The all-new Double Dare with Liza Koshy has all the trivia, physical challenges, and obstacles for the messiest game show on TV!	Paramount	0.2481
Clear and Present Danger	CIA Analyst Jack Ryan is drawn into an illegal war fought by the US government against a Colombian drug cartel.	Paramount	0.2347

Streaming Service Distribution

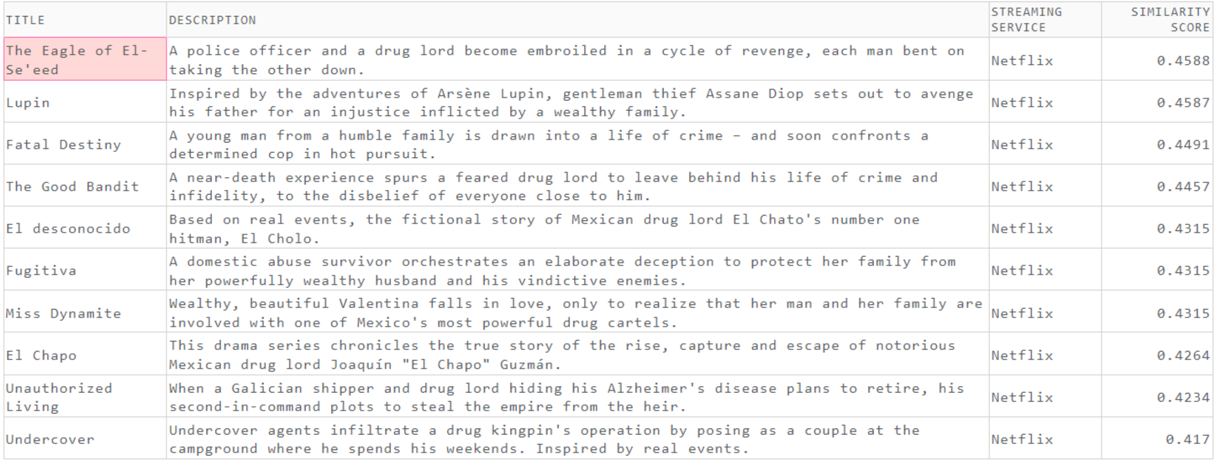


According to the output shown in the above screenshots, it seems that most of the shows the consumer would potentially watch would originate from the Paramount streaming service. This is so as nine of the recommended titles are part of the Paramount title catalog whilst only one of the recommended titles is from Disney Plus. Interestingly, the recommended title with the highest similarity score is “iCarly”. Now, the genres and rating of the recommended title differs greatly to that of the selected title “Ganglands”; this is representative to the similarity score. Although, both TV shows are Dramas. The interesting aspect of both shows is that it involves elements of family that is integral component in driving the plot forward. It should be noted that the similarity scores of all the recommended titles are below 0.5. Thus, if the consumer is feeling experimental, they can consider subscribing to the Paramount & Disney Plus streaming services. If they are not feeling experimental, then they should not spend money on subscribing.

Scenario 4

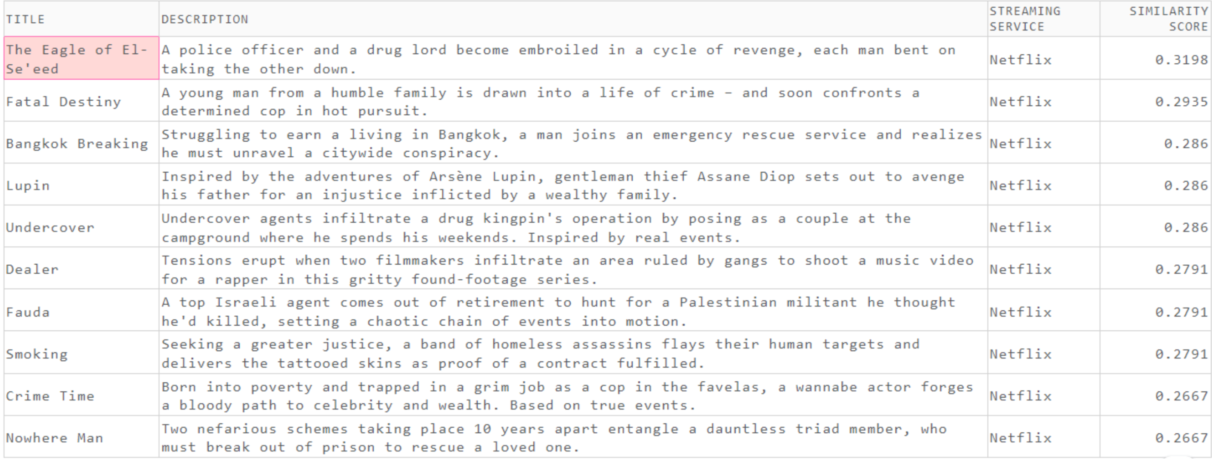
Let us say the consumer wants to compare the quality of recommendations of TF-IDF to that of NGrams (Unigrams, Bigrams, Trigrams) as the options are available. The selected title is still “Ganglands”.

Unigrams Recommendation Results



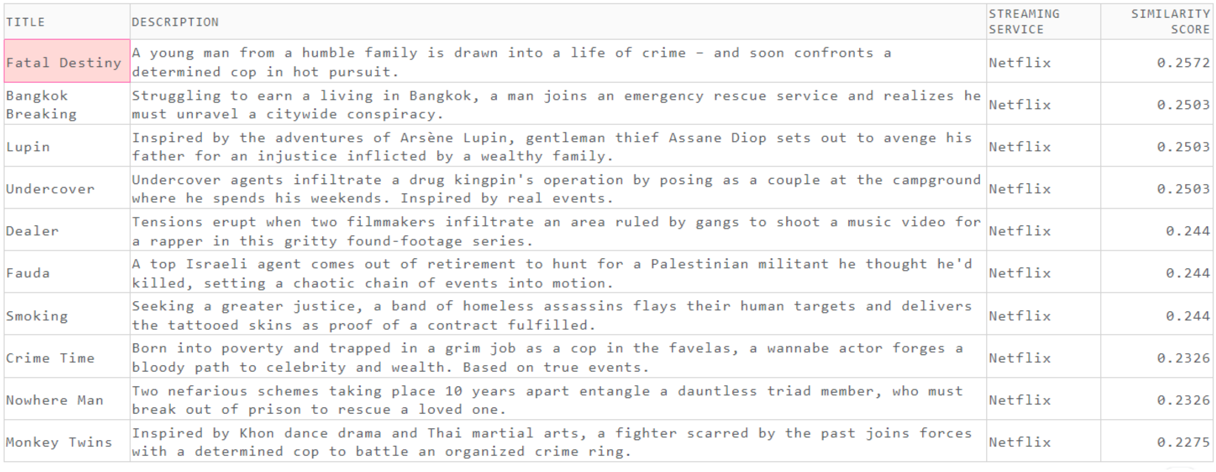
TITLE	DESCRIPTION	STREAMING SERVICE	SIMILARITY SCORE
The Eagle of El-Se'eed	A police officer and a drug lord become embroiled in a cycle of revenge, each man bent on taking the other down.	Netflix	0.4588
Lupin	Inspired by the adventures of Arsène Lupin, gentleman thief Assane Diop sets out to avenge his father for an injustice inflicted by a wealthy family.	Netflix	0.4587
Fatal Destiny	A young man from a humble family is drawn into a life of crime - and soon confronts a determined cop in hot pursuit.	Netflix	0.4491
The Good Bandit	A near-death experience spurs a feared drug lord to leave behind his life of crime and infidelity, to the disbelief of everyone close to him.	Netflix	0.4457
El desconocido	Based on real events, the fictional story of Mexican drug lord El Chato's number one hitman, El Cholo.	Netflix	0.4315
Fugitiva	A domestic abuse survivor orchestrates an elaborate deception to protect her family from her powerfully wealthy husband and his vindictive enemies.	Netflix	0.4315
Miss Dynamite	Wealthy, beautiful Valentina falls in love, only to realize that her man and her family are involved with one of Mexico's most powerful drug cartels.	Netflix	0.4315
El Chapo	This drama series chronicles the true story of the rise, capture and escape of notorious Mexican drug lord Joaquín "El Chapo" Guzmán.	Netflix	0.4264
Unauthorized Living	When a Galician shipper and drug lord hiding his Alzheimer's disease plans to retire, his second-in-command plots to steal the empire from the heir.	Netflix	0.4234
Undercover	Undercover agents infiltrate a drug kingpin's operation by posing as a couple at the campground where he spends his weekends. Inspired by real events.	Netflix	0.417

Bigrams Recommendation Results



TITLE	DESCRIPTION	STREAMING SERVICE	SIMILARITY SCORE
The Eagle of El-Se'eed	A police officer and a drug lord become embroiled in a cycle of revenge, each man bent on taking the other down.	Netflix	0.3198
Fatal Destiny	A young man from a humble family is drawn into a life of crime - and soon confronts a determined cop in hot pursuit.	Netflix	0.2935
Bangkok Breaking	Struggling to earn a living in Bangkok, a man joins an emergency rescue service and realizes he must unravel a citywide conspiracy.	Netflix	0.286
Lupin	Inspired by the adventures of Arsène Lupin, gentleman thief Assane Diop sets out to avenge his father for an injustice inflicted by a wealthy family.	Netflix	0.286
Undercover	Undercover agents infiltrate a drug kingpin's operation by posing as a couple at the campground where he spends his weekends. Inspired by real events.	Netflix	0.286
Dealer	Tensions erupt when two filmmakers infiltrate an area ruled by gangs to shoot a music video for a rapper in this gritty found-footage series.	Netflix	0.2791
Fauda	A top Israeli agent comes out of retirement to hunt for a Palestinian militant he thought he'd killed, setting a chaotic chain of events into motion.	Netflix	0.2791
Smoking	Seeking a greater justice, a band of homeless assassins flays their human targets and delivers the tattooed skins as proof of a contract fulfilled.	Netflix	0.2791
Crime Time	Born into poverty and trapped in a grim job as a cop in the favelas, a wannabe actor forges a bloody path to celebrity and wealth. Based on true events.	Netflix	0.2667
Nowhere Man	Two nefarious schemes taking place 10 years apart entangle a dauntless triad member, who must break out of prison to rescue a loved one.	Netflix	0.2667

Trigrams Recommendation Results



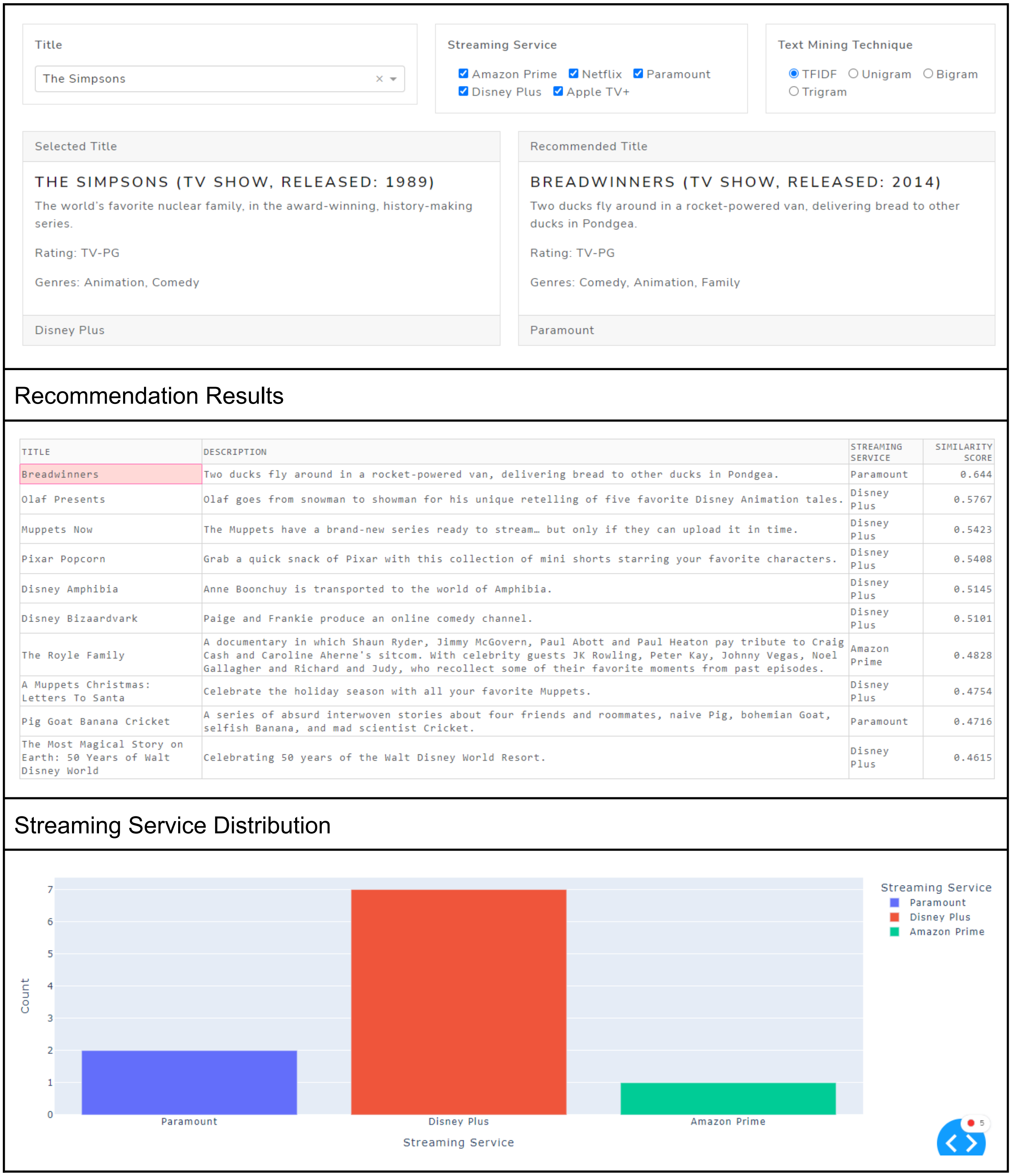
TITLE	DESCRIPTION	STREAMING SERVICE	SIMILARITY SCORE
Fatal Destiny	A young man from a humble family is drawn into a life of crime – and soon confronts a determined cop in hot pursuit.	Netflix	0.2572
Bangkok Breaking	Struggling to earn a living in Bangkok, a man joins an emergency rescue service and realizes he must unravel a citywide conspiracy.	Netflix	0.2503
Lupin	Inspired by the adventures of Arsène Lupin, gentleman thief Assane Diop sets out to avenge his father for an injustice inflicted by a wealthy family.	Netflix	0.2503
Undercover	Undercover agents infiltrate a drug kingpin's operation by posing as a couple at the campground where he spends his weekends. Inspired by real events.	Netflix	0.2503
Dealer	Tensions erupt when two filmmakers infiltrate an area ruled by gangs to shoot a music video for a rapper in this gritty found-footage series.	Netflix	0.244
Fauda	A top Israeli agent comes out of retirement to hunt for a Palestinian militant he thought he'd killed, setting a chaotic chain of events into motion.	Netflix	0.244
Smoking	Seeking a greater justice, a band of homeless assassins flays their human targets and delivers the tattooed skins as proof of a contract fulfilled.	Netflix	0.244
Crime Time	Born into poverty and trapped in a grim job as a cop in the favelas, a wannabe actor forges a bloody path to celebrity and wealth. Based on true events.	Netflix	0.2326
Nowhere Man	Two nefarious schemes taking place 10 years apart entangle a dauntless triad member, who must break out of prison to rescue a loved one.	Netflix	0.2326
Monkey Twins	Inspired by Khon dance drama and Thai martial arts, a fighter scarred by the past joins forces with a determined cop to battle an organized crime ring.	Netflix	0.2275

Interestingly, the results for Unigrams and Bigrams are somewhat similar such that the recommended title with the highest similarity score are the same (i.e., “The Eagle of El-Se’eed”), although the numerical similarity score itself differs greatly. There is a certain degree of variability of subsequent recommended titles in the Unigrams results to that of Bigrams results. For example, the titles “Lupin” and “Fatal Destiny” are second and third, respectively, in the Unigrams recommendation results. Although, “Lupin” and “Fatal Destiny” are fourth and second in the recommendation results table, respectively, in the Bigrams recommendation results. There are titles such as “El Chapo” in the Unigram recommendation results, that do not appear in the Bigram recommendation results. Interestingly, the top recommended title from the Unigrams and Bigrams recommendation results does not appear at all in the recommendation results of Trigrams. Instead, “Fatal Destiny” is the top recommended title when using Trigrams, although the similarity score is even lower than that of both Unigrams and Trigrams. Also note that the top recommended title that appeared in the TF-IDF recommendation results, “Narcos”, does not even appear in all of the NGrams recommendation results.

Scenario 5

The consumer is now exhausted of watching serious shows involving violence such as “Ganglands” and similar shows. They enjoy watching “The Simpsons”, which is a long-time classic and has been on-the-air since 1989. They would like to find similar shows that are available on all the streaming services, and will use the TF-IDF technique again.

Selected Title vs Recommended Title with Highest Similarity



Title

The Simpsons ×

Streaming Service

☑ Amazon Prime ☑ Netflix ☑ Paramount
☑ Disney Plus ☑ Apple TV+

Text Mining Technique

● TFIDF ○ Unigram ○ Bigram
○ Trigram

Selected Title

THE SIMPSONS (TV SHOW, RELEASED: 1989)

The world's favorite nuclear family, in the award-winning, history-making series.

Rating: TV-PG

Genres: Animation, Comedy

Disney Plus

Recommended Title

BREADWINNERS (TV SHOW, RELEASED: 2014)

Two ducks fly around in a rocket-powered van, delivering bread to other ducks in Pondgea.

Rating: TV-PG

Genres: Comedy, Animation, Family

Paramount

Recommendation Results

TITLE	DESCRIPTION	STREAMING SERVICE	SIMILARITY SCORE
Breadwinners	Two ducks fly around in a rocket-powered van, delivering bread to other ducks in Pondgea.	Paramount	0.644
Olaf Presents	Olaf goes from snowman to showman for his unique retelling of five favorite Disney Animation tales.	Disney Plus	0.5767
Muppets Now	The Muppets have a brand-new series ready to stream… but only if they can upload it in time.	Disney Plus	0.5423
Pixar Popcorn	Grab a quick snack of Pixar with this collection of mini shorts starring your favorite characters.	Disney Plus	0.5408
Disney Amphibia	Anne Boonchuy is transported to the world of Amphibia.	Disney Plus	0.5145
Disney Bizaardvark	Paige and Frankie produce an online comedy channel.	Disney Plus	0.5101
The Royle Family	A documentary in which Shaun Ryder, Jimmy McGovern, Paul Abott and Paul Heaton pay tribute to Craig Cash and Caroline Aherne's sitcom. With celebrity guests JK Rowling, Peter Kay, Johnny Vegas, Noel Gallagher and Richard and Judy, who recollect some of their favorite moments from past episodes.	Amazon Prime	0.4828
A Muppets Christmas: Letters To Santa	Celebrate the holiday season with all your favorite Muppets.	Disney Plus	0.4754
Pig Goat Banana Cricket	A series of absurd interwoven stories about four friends and roommates, naive Pig, bohemian Goat, selfish Banana, and mad scientist Cricket.	Paramount	0.4716
The Most Magical Story on Earth: 50 Years of Walt Disney World	Celebrating 50 years of the Walt Disney World Resort.	Disney Plus	0.4615

Streaming Service Distribution

As can be seen from the above screenshots of the results, there is a higher distribution of streaming services the user could consider subscribing to based on the selected title. Majority of the recommended titles come from Disney Plus, although the recommended title with the highest similarity score, “Breadwinners”, is actually available on the Paramount streaming service. The recommendation makes sense as the premise of both selected title and recommended title are absurd in nature, with PG rating and near identical genres. This is further supported by the fact that the similarity score of the top recommended title is above 0.5, which signifies a strong similarity to the selected title. As the similarity scores of the subsequent recommended titles is relatively high, the consumer should definitely consider subscribing to Paramount and Disney Plus especially. The unique observation regarding the results shown above is that the Netflix streaming service does not appear at all despite a high proportion of the final unified dataset containing Netflix titles. Also, TV Shows such as Family Guy and Bob's Burgers, which were inspired by "The Simpsons", do not appear at all in the recommendation results.

Conclusion

- The TF-IDF technique yielded more overall recommendation results that had similarity scores greater than 0.5, compared to NGrams
- There could be more distribution of streaming services recommended for streaming services when not considering the Netflix streaming service in the recommendation system
- NGrams may not yield conclusive results when NOT considering the importance of words
- More generic or popular long-standing shows could output a more varied distribution of recommended streaming services
- Due to the large title catalog of Netflix, when not considering the streaming service in the recommendation system, there may be less recommended titles that are above 0.5, regardless of the textual analysis used

Reflection

If we were to attempt this project again, we would consider more than one data source rather than solely Kaggle. Utilizing APIs, where we can extract detailed plot description data. There may be API request limits in place which would limit us from extracting the data that we need all in one go, and so we would have to incrementally extract the data over a period of time. Paid APIs may not have low request limits so we could extract all in one go. The front-end application could be further improved to provide more features so that the user can manipulate the parameters utilized by the 'TFIDFVectorizer' and 'CountVectorizer' functions. This way, the user can fine-tune their recommendation results to their liking and try to optimize the parameters of the functions to get more accurate results. This makes the application even more data analytics oriented than it already is, as well as educational as a student as one is exposed to the techniques by simply utilizing the application itself. Another feature we would have liked to incorporate is to display a visualization of the corpus, logarithmic chart of TF-IDF, and term frequency chart. Although, due to how slow the application runs already, we decided to not pursue it to avoid further lag due to the fact that the project requirement is to show a live demonstration during the presentation day. Thus, the application would not run optimally if so many features are bloated into it with several instances of unsupervised learning taking place.

References

Chouinard, J. (2022). How To Build A Recommender System With TF-IDF And NMF (Python). Search Engine Journal. Retrieved 26 July 2022, from https://www.searchenginejournal.com/topic-clusters-recommender-system/436123/#:~:text=What%20Is%20TF%2DIDF%3F,frequency%20and%20inverse%20document%20frequency.

Ganesan, K. (2020). What are N-Grams? - Kavita Ganesan, PhD. Kavita Ganesan, PhD. Retrieved 26 July 2022, from https://kavita-ganesan.com/what-are-n-grams/.

Wan, M. (2020). Beginner's Guide to Building a Multi-Page Dashboard using Dash. Medium. Retrieved 24 June 2022, from https://towardsdatascience.com/beginners-guide-to-building-a-multi-page-dashboard-using-dash-5d06dbfc7599.

What Are the 4 Main Analytical Models?. FutureLearn. Retrieved 26 July 2022, from https://www.futurelearn.com/info/courses/data-analytics-for-business-basic-analysis-and-statistics/0/steps/177283.

What is a good threshold for CosineSimilarity Measure?. RapidMiner Community. (2020). Retrieved 26 July 2022, from https://community.rapidminer.com/discussion/56700/what-is-a-good-threshold-for-cosinesimilarity-measure.

Helpful Resources

- Pre-processing & TFIDF
    - https://www.youtube.com/watch?v=7WfoYI-EPtI&list=LL&index=6
    - https://www.kaggle.com/code/rushikeshdane20/build-recommendation-system-app-with-streamlit
    - https://datagy.io/python-remove-punctuation-from-string/#:~:text=One%20of%20the%20easiest%20ways,maketrans()%20method.
    - https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
    - https://appdividend.com/2020/12/16/python-string-lower-method/#:~:text=Python%20lower()%20function%20is,the%20Python%20upper()%20method.
    - https://www.analyticssteps.com/blogs/nltk-python-tutorial-beginners
    - https://swatimeena989.medium.com/beginners-guide-for-preprocessing-text-data-f3156bec85ca
    - https://www.analyticssteps.com/blogs/what-stemming-and-lemmatization-nlp
    - https://himanshulohiya.medium.com/tf-idf-d00d7ec4f362
    - https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

  - https://programminghistorian.org/en/lessons/analyzing-documents-with-tfidf#scikit-learn-settings
  - https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a
  - https://medium.com/@gianpaul.r/tokenization-and-parts-of-speech-pos-tagging-in-pythons-nltk-library-2d30f70af13b
  - https://www.dataknowsall.com/pos.html
  - https://www.geeksforgeeks.org/string-capitalize-python/
  - https://www.geeksforgeeks.org/how-to-drop-rows-in-pandas-dataframe-by-index-labels/
- Python and Dash
  - https://www.youtube.com/watch?v=-KLtU_t5bXs
  - https://dash.plotly.com/datatable/callbacks
  - https://dash.plotly.com/datatable
  - https://dash.plotly.com/dash-html-components/button
  - https://python.tutorialink.com/do-a-button-click-from-code-in-plotly-dash/
  - https://github.com/Coding-with-Adam/Dash-by-Plotly/blob/master/Callbacks/Basic%20Callback/basic_callback.py
  - https://community.plotly.com/t/how-to-wrap-text-in-cell-in-dash-table/15687/6
  - https://stackoverflow.com/questions/35164019/filter-multiple-values-using-pandas
  - https://community.plotly.com/t/datatable-active-cell-row-information-in-multiple-page-tables/45783/2
  - https://community.plotly.com/t/multiple-outputs-in-dash-now-available/19437
  - https://stackoverflow.com/questions/63592900/plotly-dash-how-to-design-the-layout-using-dash-bootstrap-components
  - https://www.statology.org/pandas-get-index-of-row/
  - http://dash-bootstrap-components.opensource.faculty.ai/docs/components/card/
  - https://dash.plotly.com/dash-core-components/graph
  - https://dash.plotly.com/sharing-data-between-callbacks
  - https://appdividend.com/2022/01/13/how-to-convert-python-list-to-tuple/#:~:text=To%20convert%20the%20Python%20list,the%20easiest%20approach%20for%20conversion.
  - https://dash.plotly.com/dash-core-components/radioitems
  - https://dash.plotly.com/datatable/style
- NGrams research
  - https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
  - https://stackabuse.com/simple-nlp-in-python-with-textblob-n-grams-detection/?ref=morioh.com&utm_source=morioh.com
  - https://www.askpython.com/python/examples/n-grams-python-nltk
  - https://www.searchenginejournal.com/topic-clusters-recommender-system/436123/
  - https://www.kdnuggets.com/2022/06/ngram-language-modeling-natural-language-processing.html
  - https://www.geeksforgeeks.org/scrape-google-ngram-viewer-using-python/

- https://www.kdnuggets.com/2018/02/recommender-engine.html
- https://financial-engineering.medium.com/justforfunpython-n-gram-to-quantify-similarity-between-sentences-2d61e68a478c
- https://datascience.stackexchange.com/questions/37073/what-would-be-the-best-way-to-map-similar-ngrams
- https://zditect.com/blog/2905059.html
- https://medium.com/fintechexplained/nlp-text-mining-algorithms-4546c6ca30a
- https://gist.github.com/gaulinmp/da5825de975ed0ea6a24186434c24fe4
- https://pythonhosted.org/ngram/ngram.html
- https://www.quora.com/What-is-the-difference-between-TfidfVectorizer-and-CountVectorizer-1
- https://thepoints.medium.com/feature-extraction-from-text-using-countvectorizer-tfidfvectorizer-9f74f38f86cc
- https://stackoverflow.com/questions/17627219/whats-the-fastest-way-in-python-to-calculate-cosine-similarity-given-sparse-mat
- https://stackoverflow.com/questions/67759152/how-to-write-a-method-that-returns-cosine-similarity-between-two-documents