Data Science

GENERIC RECOMMENDATION SYSTEM

Rutvik Deshpande, Amaan Vora, Priyal Shaha, Jahnavi Shah

Abstract

Our project involves the development of a versatile recommendation system that can provide recommendations for any dataset that adheres to a standardized format. With this model, we aim to provide a user-friendly and adaptable solution for businesses and organizations seeking to enhance their recommendation capabilities.
We plan to do some exploratory data analysis, clean the data from any anomalies and utilize the power of the categorical attributes using NLP. After that, we intend to create a clustering algorithm which would use the attributes selected by PCA to recommend Movies, TV Shows or any other data demanded by the user.

Problem

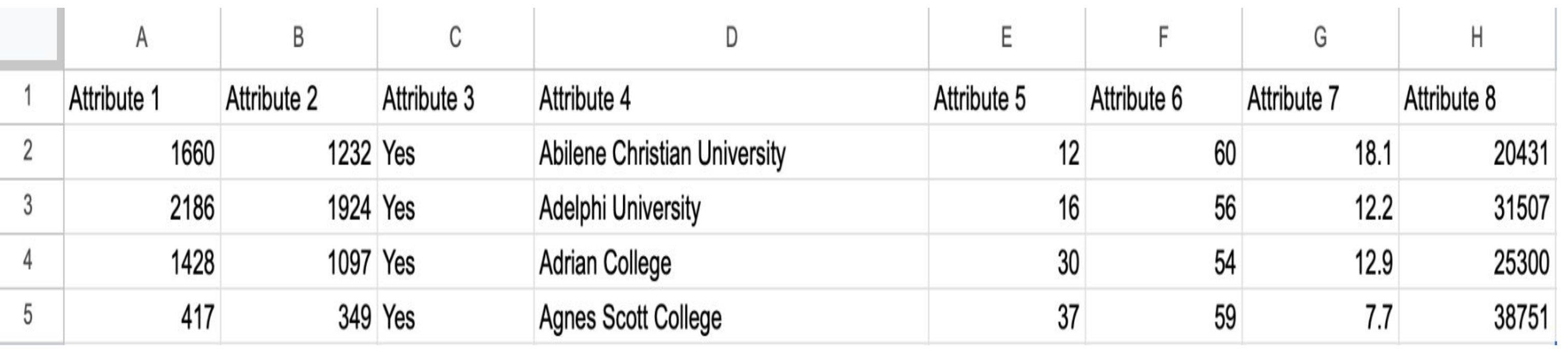
Are you tired of getting stuck in analysis paralysis? Are you indecisive?

Our recommendation system takes any data that follows *a specific format* to predict items, such as movies, books and more!
We use content based clustering algorithm along with NLP for better predictions. We have also tried using a Collaborative filtering that recommends items based on the preferences of similar users, Content based filtering that uses recommends items based on the characteristics of the items themselves and a Hybrid approach as well.

Data

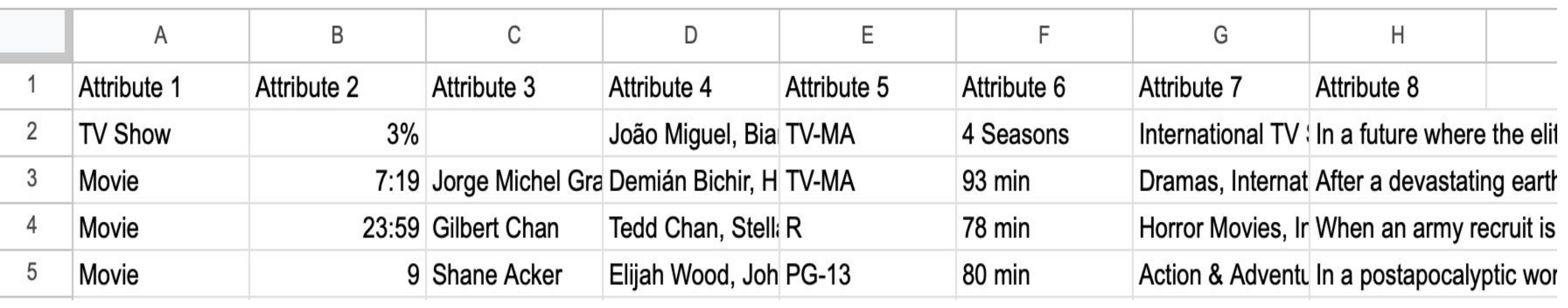
Our project uses a Netflix dataset and University information dataset.
We begin with creating a model using the Netflix data which contains 12 attributes and 7787 rows.A critical aspect of our recommendation system is data comprehension. We analyze eight attributes on both the sides, with a mixture of structured and unstructured data features. To ensure high-quality recommendations, we perform comprehensive data cleaning and feature extraction to enhance the accuracy of our data points.

COLLEGE DATA SET



	A	B	C	D	E	F	G	H
1	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	Attribute 7	Attribute 8
2	1660	1232	Yes	Abilene Christian University	12	60	18.1	20431
3	2186	1924	Yes	Adelphi University	16	56	12.2	31507
4	1428	1097	Yes	Adrian College	30	54	12.9	25300
5	417	349	Yes	Agnes Scott College	37	59	7.7	38751

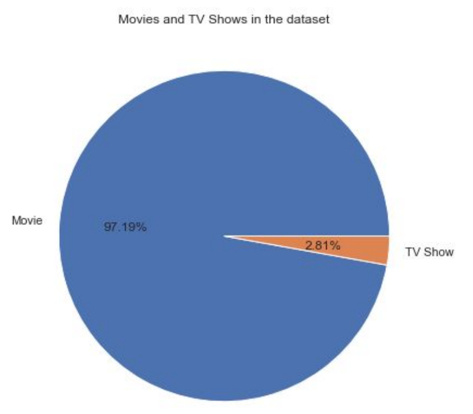
NETFLIX DATA SET



	A	B	C	D	E	F	G	H
1	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	Attribute 7	Attribute 8
2	TV Show	3%		João Miguel, Bia	TV-MA	4 Seasons	International TV :	In a future where the elit
3	Movie	7:19	Jorge Michel Gra	Demián Bichir, H	TV-MA	93 min	Dramas, Internat	After a devastating eartl
4	Movie	23:59	Gilbert Chan	Tedd Chan, Stell:	R	78 min	Horror Movies, Ir	When an army recruit is
5	Movie	9	Shane Acker	Elijah Wood, Joh	PG-13	80 min	Action & Adventu	In a postapocalyptic wor

Methodology

Data Preparation and EDA



We start off by dividing the data into individual components.

Multivariate Exploratory Data Analysis gave us an understanding of further evaluations needed.

Visual representations of our input data include plots of entertainment media distribution over time, length of media, and a word cloud depicting common entries.

family
life
find
man
love
take
two
new
woman
young
friend
father
mother
teen
son
home

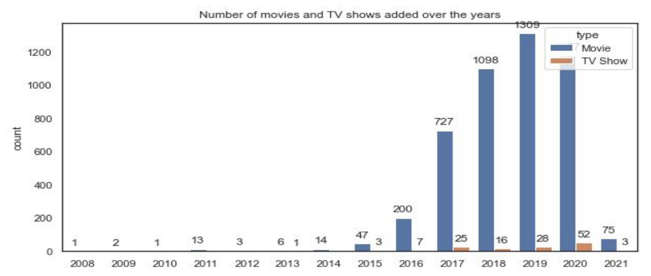
We tokenized, removed stopwords and lemmatized data which helped us retain the atomicity of our data points for our clustering algorithms.

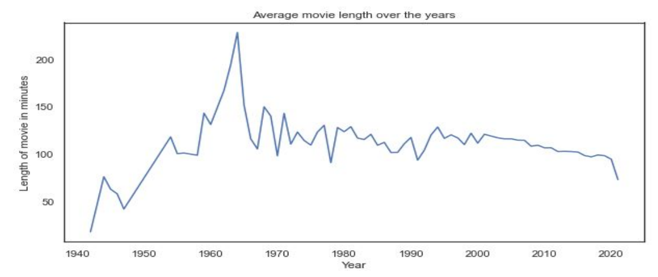
Finally, we divided up our data into clustering attributes which would be essential to our algorithm.

$$TF = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

$$Idata(t) = log_e(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it})$$

$$TFIdata = TF * Idata$$

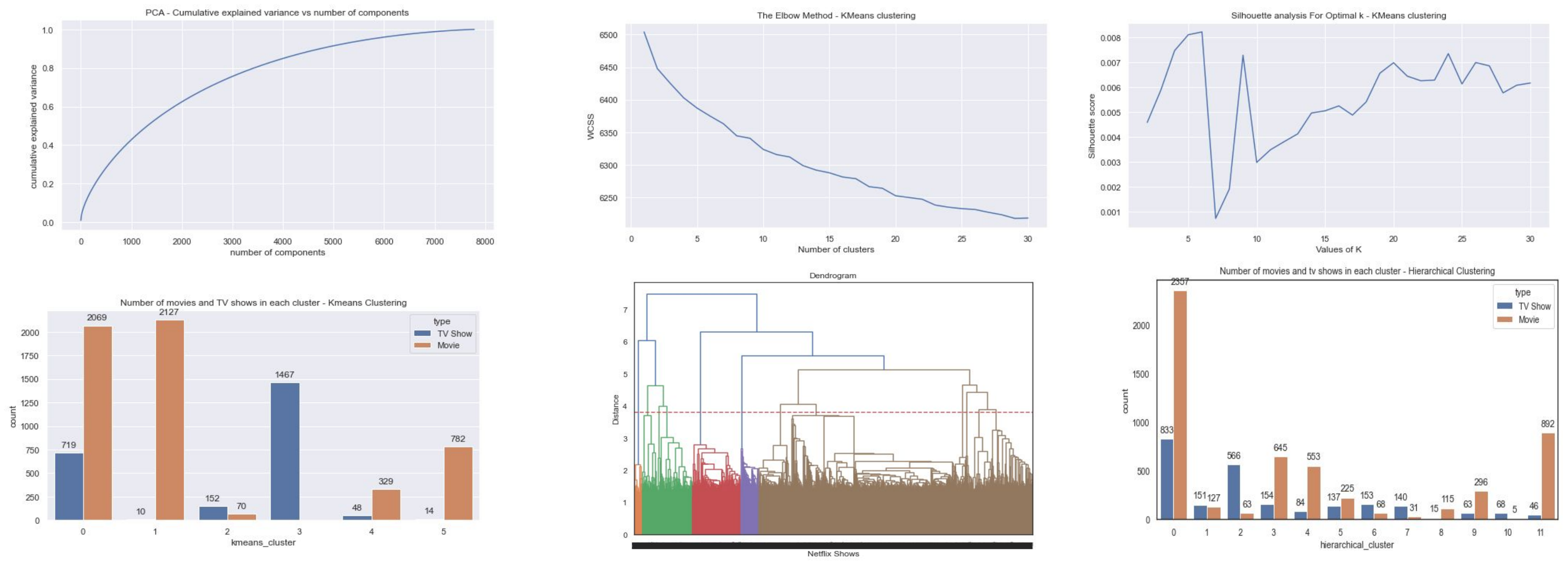




Model

Algorithm: K means Clustering

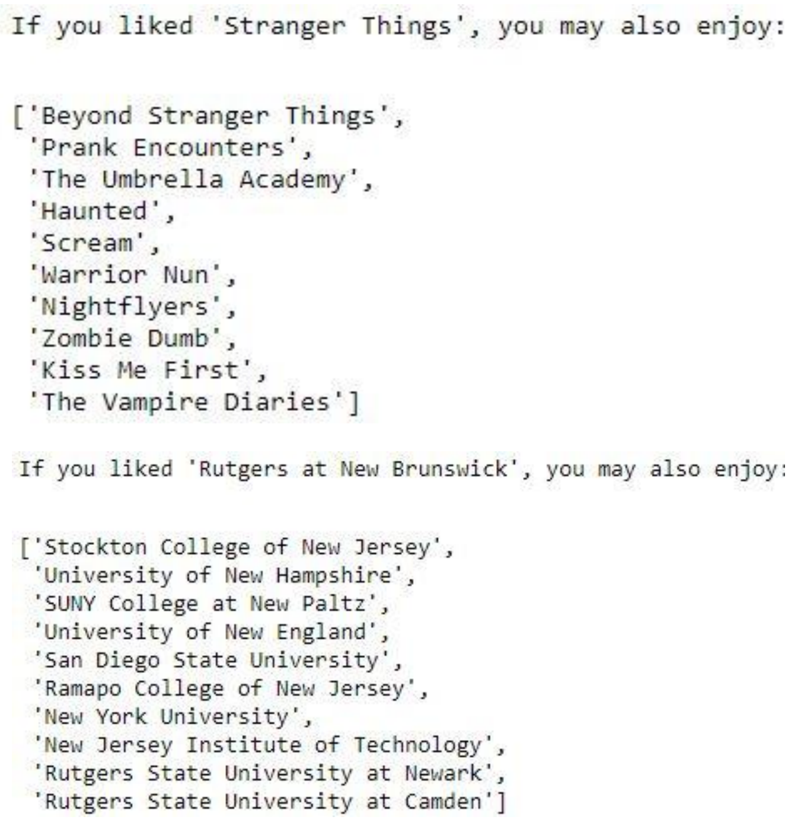
1. Utilizing Principal component analysis we identified that the cumulative variance in our data plateaus at around 4000 values. Therefore, we reduce our attributes to 4000.
2. We used WCSS by plotting an elbow method and a silhouette analysis, showing the optimal clusters to be 6. For further improvement, we employed hierarchical clustering.
3. We visualize a dendrogram and determine the number of optimal clusters to be 12 at a distance of 3.8 units.
4. We ran similar testing for college data and observed identical results.
5. Our recommendation system utilizes a cosine function to determine contextual similarity, providing a generic and adaptable solution.
6. This function is calculated using a dot product of two vectors divided by a magnitude value.
7. We perform count vectorization to flatten values for our cosine imputation.



Findings and Evaluation

1. From our recommender system, we were able to extract titles of entertainment media and universities, depending upon our dataset.
2. Testing against manual recommendations, we found it to give an output based on certain attributes like description and tuition fee more than others.
3. We were also able to generate a Directed Acyclic Graph using a NetworkX layer that helped visualize our clustering algorithm.
4. Overall, hierarchical clustering helped improve our centroids, which further reduced distances of datapoints, improving our algorithm.

$$Cos(\theta) = \frac{A\ .\ B}{|A|\ .\ |B|}$$

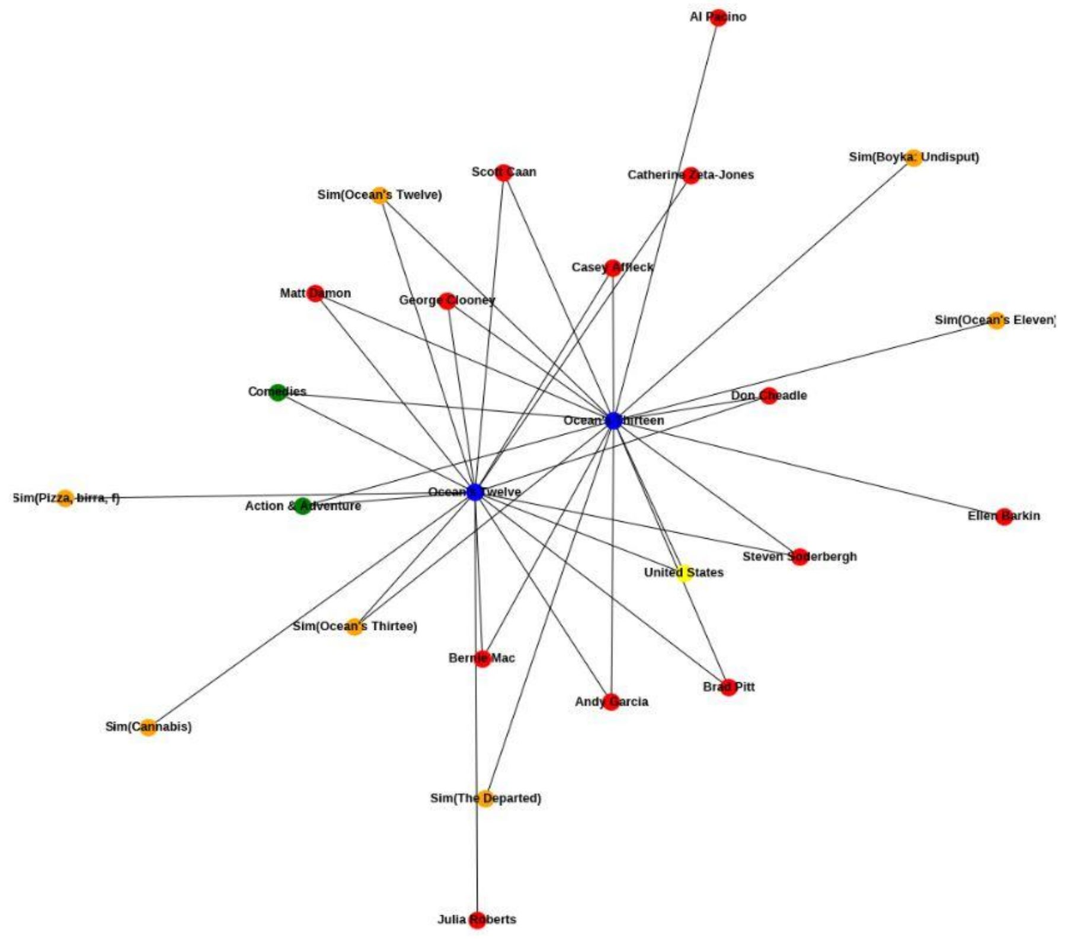


```
If you liked 'Stranger Things', you may also enjoy:

['Beyond Stranger Things',
 'Prank Encounters',
 'The Umbrella Academy',
 'Haunted',
 'Scream',
 'Warrior Nun',
 'Nightflyers',
 'Zombie Dumb',
 'Kiss Me First',
 'The Vampire Diaries']

If you liked 'Rutgers at New Brunswick', you may also enjoy:

['Stockton College of New Jersey',
 'University of New Hampshire',
 'SUNY College at New Paltz',
 'University of New England',
 'San Diego State University',
 'Ramapo College of New Jersey',
 'New York University',
 'New Jersey Institute of Technology',
 'Rutgers State University at Newark',
 'Rutgers State University at Camden']
```



Conclusions & Future Scope

1. The objective of this project was to build a generic recommender system that would be able to take in any data points within a regularized format and produce a clustered output.
2. The nature of a regularized format depended upon the tabular structure of our data, and so we were tied down by its drawbacks.
3. Another drawback was the format of our data, which resulted in some generic results which could be avoided with stronger, more trained algorithms.
4. We hope to explore a methodology with transformers like GPT and BERT that would be able to take in many more forms of unstructured data in order to produce better results.
5. Additionally, we also plan on furthering this project by opting for all datasets being provided at once for more training points.
6. This would enable us to use better feature selection and filtering techniques to accelerate more accurate results.

FSRM 588 Financial Data Mining Poster Presentation. Spring 2023. Rutgers University.