Movie Recommendation system

Introduction:

In the dynamic landscape of the digital era, an overwhelming influx of information and choices confronts users at every turn. Whether navigating e-commerce platforms, exploring online resources, or engaging with streaming content, the challenge of pinpointing relevant and captivating options has evolved into a pervasive concern. Enter recommendation systems – the pivotal driving forces shaping personalized experiences and revolutionizing digital content interactions. Fueled by avant-garde technologies, intricate algorithms, and astute data analytics, these systems transcend mere suggestion; they possess the capability to not only discern, but also anticipate and cater to individual predilections with unparalleled precision. By meticulously deciphering user behavioral patterns, historical trends, and contextual nuances, recommendation systems unlock a boundless realm of possibilities that redefine the very essence of convenience, engagement, and user satisfaction.

The Significance of Recommendation Systems:

In the ever-expanding digital realm, recommendation systems emerge as indispensable guides, seamlessly enhancing online experiences with tailored precision. Beyond the convenience they afford users, these systems wield profound significance for businesses. Serving as catalysts for increased sales and enhanced customer engagement, they wield the power to not only suggest potential purchases, but also provide invaluable insights into consumer preferences, consequently fostering the refinement of products and services. The intricate dance between user preferences and algorithmic ingenuity transforms recommendation systems into the digital envoys that enrich our online lives. These systems serve as the compass guiding us to captivating movies, facilitating prudent purchasing decisions, and nurturing exploration, while concurrently furnishing enterprises with the competitive edge necessary for success in the modern market milieu.

Our Movie Recommendation System:

In today's digital epoch, the proliferation of cinematic choices can overwhelm individuals seeking films aligned with their unique tastes. Enter movie recommendation systems – a bespoke solution to this contemporary conundrum, meticulously engineered to provide personalized cinematic guidance and redefine the very contours of cinematic exploration. These cerebral constructs leverage the intricate interplay of user behavior and movie attributes to craft customized movie suggestions, thereby elevating user experiences and immersing them in a realm of cinematic marvels.

With an unwavering commitment to harnessing the potency of Artificial Intelligence solutions, we embarked on the creation of a movie recommendations chatbot. Our endeavor sought to address a global challenge – the optimization of AI in domains like entertainment recommendations. Central to our mission was the core challenge of predicting and suggesting movies poised to resonate with individual users, predicated upon their past interactions and movie attributes. That’s why we proposed to create a movie recommendations chatbot to deal with a challenge that faces the world which is providing AI solutions and get the most benefit out of it in fields like entertainment suggestions.

The core challenge addressed by a movie recommendation system is to predict and suggest movies that a user is likely to enjoy, based on their past interactions and the characteristics of movies.

Concepts and Challenges:

Our pursuit of implementing an effective movie recommendation system brought us face to face with a triad of pivotal concepts and challenges that underpin its development. These foundational aspects encapsulate the essence of our endeavor, each presenting a unique sphere of consideration.

1. Data Acquisition and Selection: The Quest for an Apt Dataset

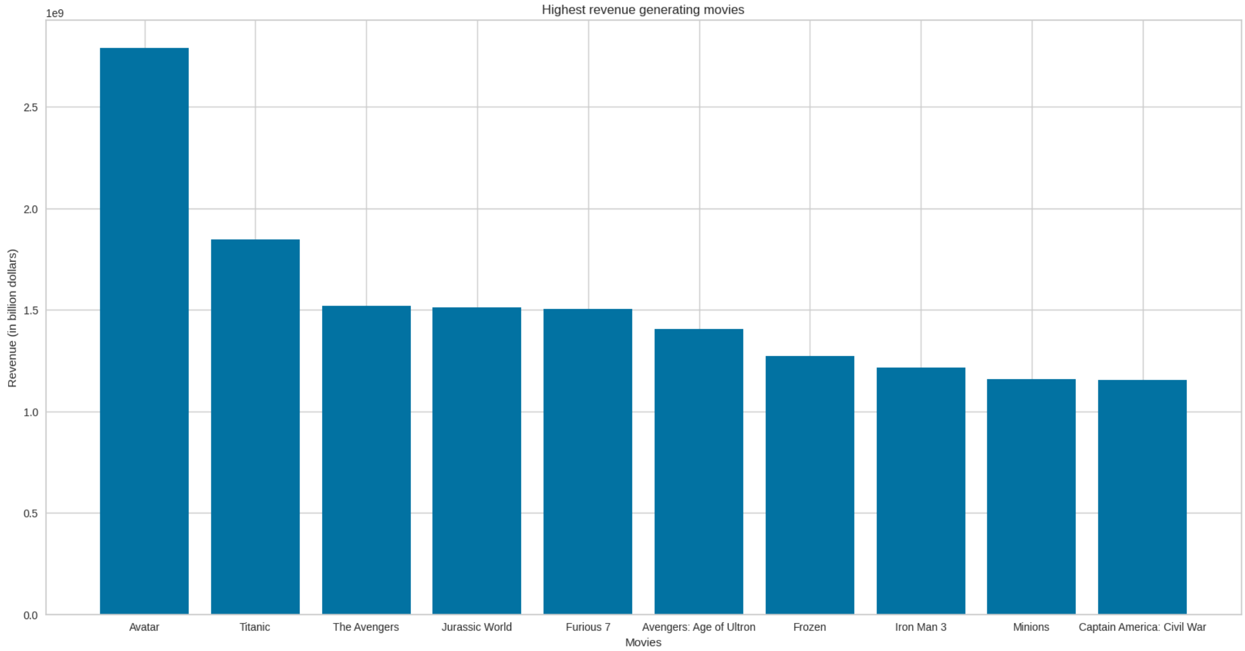
At the heart of our journey lay the critical task of acquiring a dataset that aligns seamlessly with our system's requirements and dependencies. This dataset forms the bedrock upon which our recommendation system hinges, as it serves as the wellspring from which the system draws insights to craft tailored movie suggestions.

Dataset:

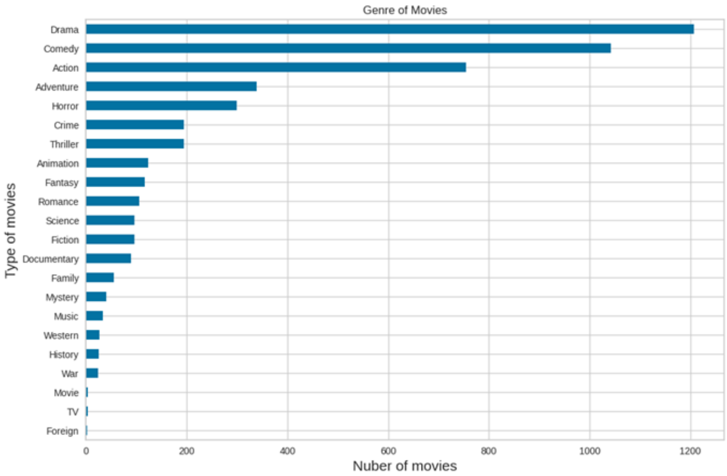
In devising our recommendation system, we homed in on four pivotal attributes - genre, keywords, cast, and director - as the key determinants for precise movie recommendations. To cater to these requirements, we meticulously evaluated datasets and found resonance with the TMDB movie dataset 5000. This comprehensive repository boasts 5000 distinct movies spanning the years 1910 to 2015, encompassing a multitude of attributes such as 'budget,' 'genres,' 'keywords,' 'cast,' 'director,' and more. This judicious selection endowed us with a robust foundation to craft our recommendation system.

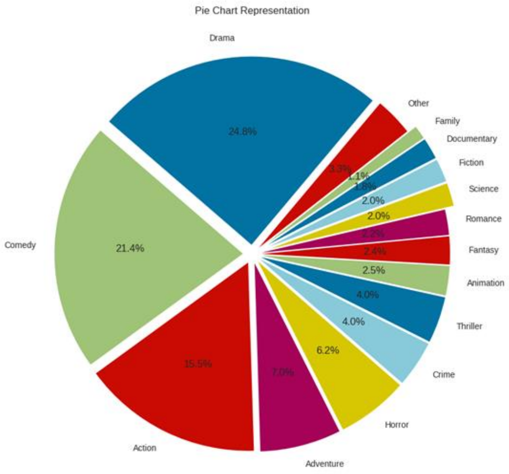
Those are some visuals for the Exploratory data analysis:

10 movies with most revenue

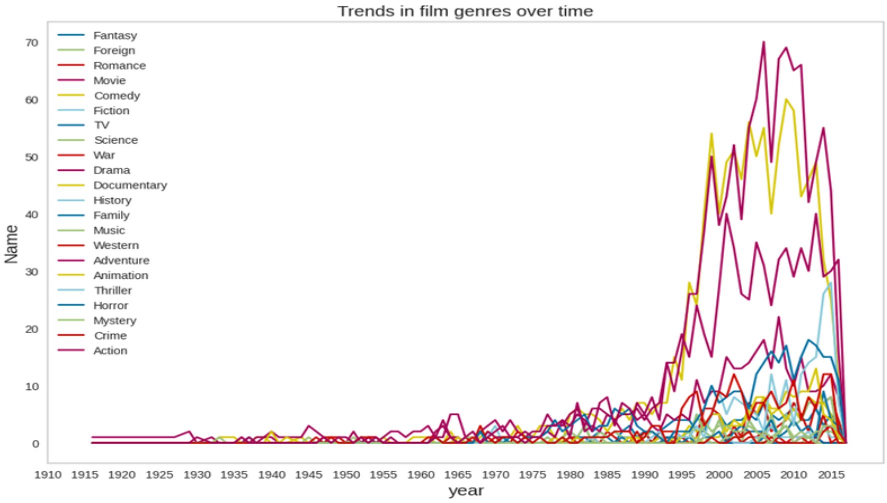


Total movies in each Genre

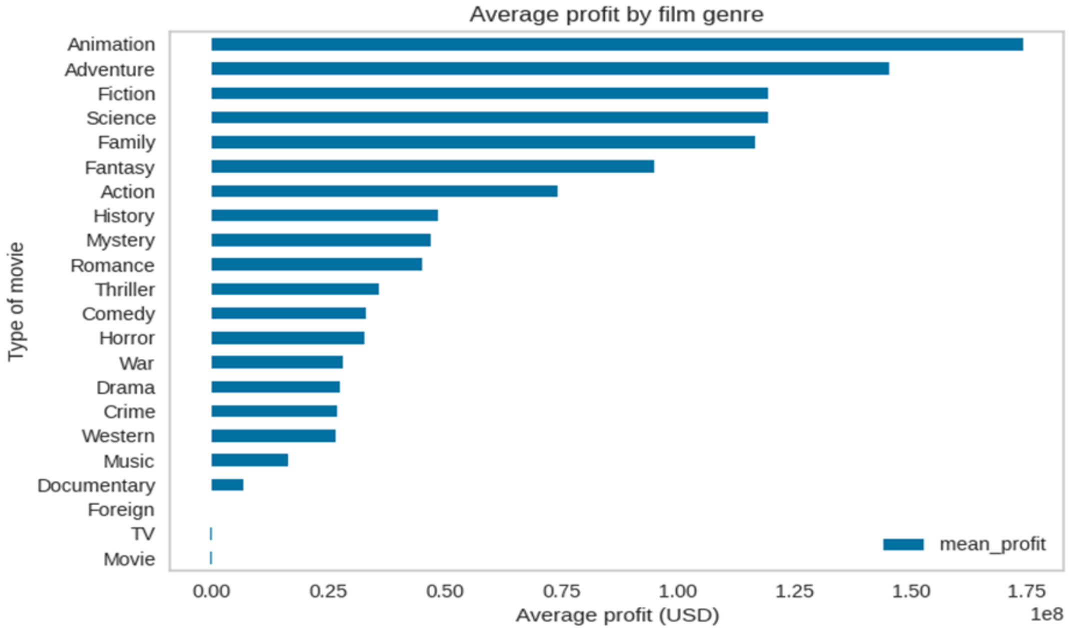




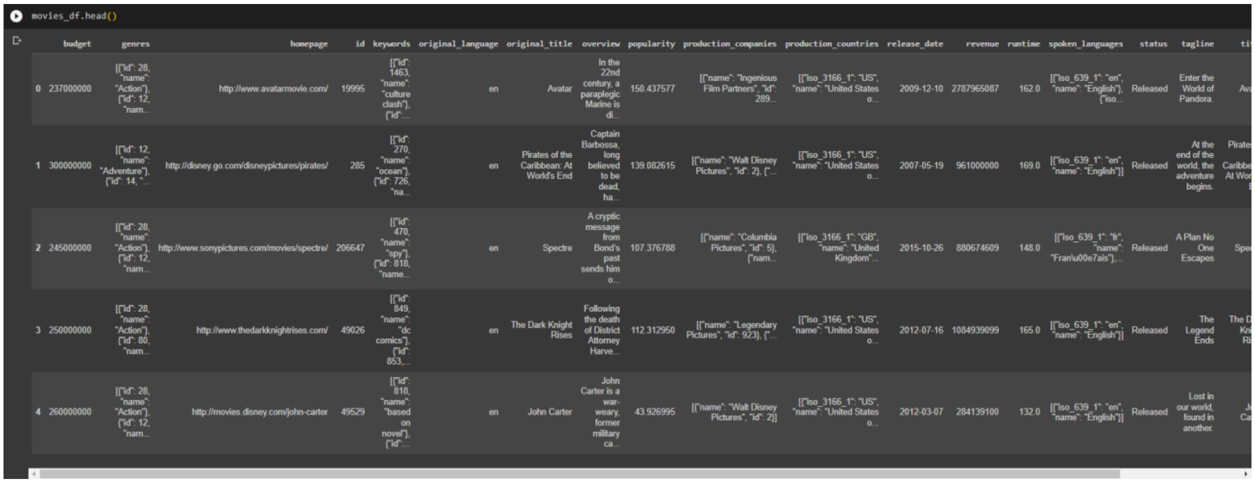
Year-wise comparison of different genres:



Average profit by film genre

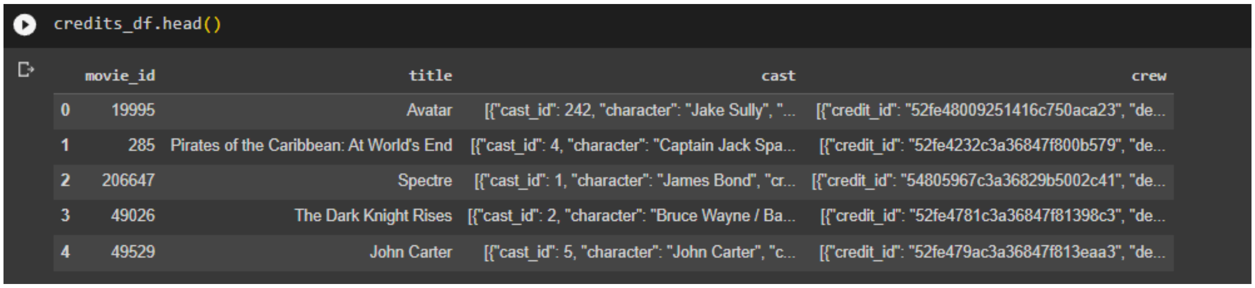


Here is some raw screenshots of the movies data since it is divided into two csv files :



```
movies_df.head()
```

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_companies	production_countries	release_date	revenue	runtime	spoken_languages	status	tagline	ti
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar	In the 22nd century, a paraplegic Marine is di...	150.437577	[{"name": "Ingenious Film Partners", "id": 289...	[{"iso_3166_1": "US", "name": "United States o...	2009-12-10	2787965087	162.0	[{"iso_639_1": "en", "name": "English"}, {"iso...	Released	Enter the World of Pandora.	Ava
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	en	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	139.082615	[{"name": "Walt Disney Pictures", "id": 2}, {"...	[{"iso_3166_1": "US", "name": "United States o...	2007-05-19	961000000	169.0	[{"iso_639_1": "en", "name": "English"}]	Released	At the end of the world, the adventure begins.	Pirates Caribbe At Wor E
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.sonypictures.com/movies/spectre/	206647	[{"id": 470, "name": "spy"}, {"id": 818, "name...	en	Spectre	A cryptic message from Bond's past sends him o...	107.376788	[{"name": "Columbia Pictures", "id": 5}, {"nam...	[{"iso_3166_1": "GB", "name": "United Kingdom"...	2015-10-26	880674609	148.0	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"},...	Released	A Plan No One Escapes	Spe
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	http://www.thedarkknightrises.com/	49026	[{"id": 849, "name": "dc comics"}, {"id": 853,...	en	The Dark Knight Rises	Following the death of District Attorney Harve...	112.312950	[{"name": "Legendary Pictures", "id": 923}, {"...	[{"iso_3166_1": "US", "name": "United States o...	2012-07-16	1084939099	165.0	[{"iso_639_1": "en", "name": "English"}]	Released	The Legend Ends	The D Kni Ri
4	260000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://movies.disney.com/john-carter	49529	[{"id": 818, "name": "based on novel"}, {"id":...	en	John Carter	John Carter is a war-weary, former military ca...	43.926995	[{"name": "Walt Disney Pictures", "id": 2}]	[{"iso_3166_1": "US", "name": "United States o...	2012-03-07	284139100	132.0	[{"iso_639_1": "en", "name": "English"}]	Released	Lost in our world, found in another.	J Ca



```
credits_df.head()
```

	movie_id	title	cast	crew
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	49026	The Dark Knight Rises	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	49529	John Carter	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

Preprocessing: Sculpting Raw Data into Refined Insights

With our dataset in hand, the journey ventured into the realm of preprocessing - a transformative phase that molds raw data into a refined, harmonious whole. Here, a sequence of strategic steps unfolded, each contributing to the creation of an organized and streamlined dataset poised for further analysis and algorithmic application.

Key Preprocessing Steps:

Integration of Data Sources: The dataset was an amalgamation of distinct CSV files, demarcated into credits and movies data. We embarked on a merger, harmonizing these disparate sources into a unified repository to enhance data coherence.

Feature Extraction and Refinement: The potency of the data lay in its attributes. We meticulously extracted pertinent details, harnessing the 'crew' feature to distill the director's identity and unpacking the 'cast' feature to unveil the ensemble cast members.

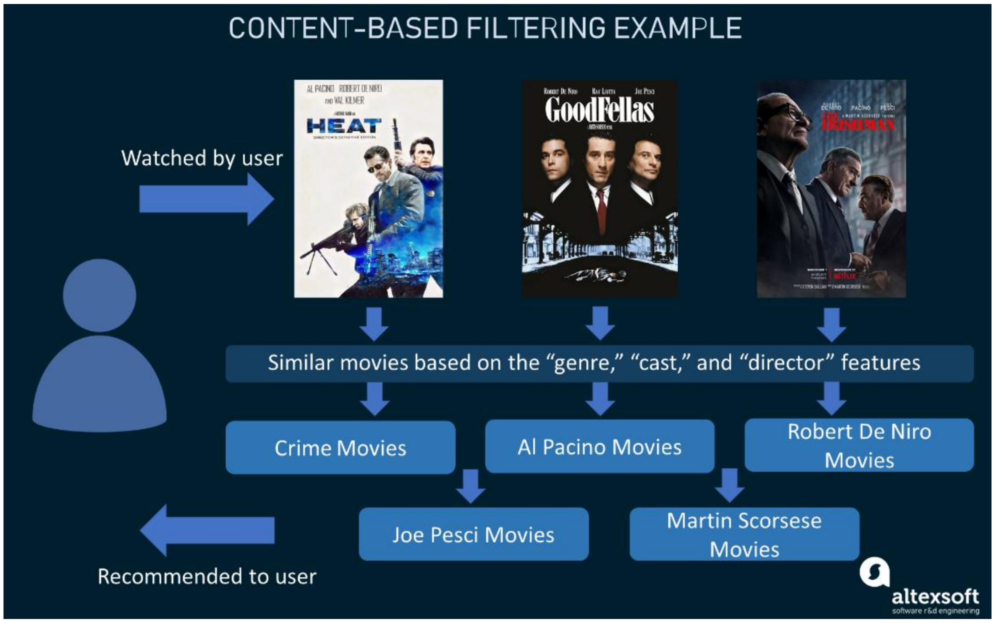
Data Cleansing and Standardization: A pivotal endeavor involved purging the dataset of impediments such as stopwords, irregular inputs, and inconsistencies. Further, we unified the input format by transforming it into lowercase, ensuring a seamless and consistent analysis process.

Feature Fusion for Enriched Insight: To amplify the potency of our dataset, we orchestrated the harmonious union of genre, cast, director, and keywords into a singular feature. This conglomerate, primed for vectorization, encapsulated the holistic essence of movie attributes.

Vectorization for Algorithmic Application: With our refined dataset in place, we applied vectorization techniques, bestowing numerical representation upon textual attributes. This pivotal step set the stage for the subsequent algorithmic prowess of our recommendation system.

Our journey thus far has laid the groundwork for the development of a robust movie recommendation system. Through meticulous dataset selection, insightful preprocessing, and strategic feature amalgamation, we have carved a path towards a system poised to illuminate personalized cinematic journeys for users.

2-Content-Based Movie Recommendation with Cosine Similarity:



CONTENT-BASED FILTERING EXAMPLE

Watched by user

Similar movies based on the "genre," "cast," and "director" features

Crime Movies

Al Pacino Movies

Robert De Niro Movies

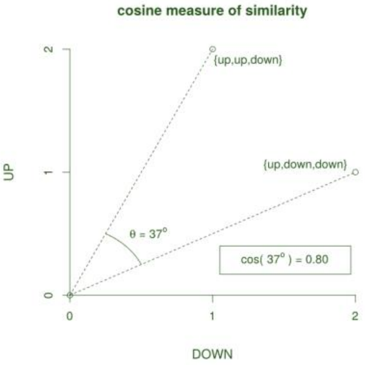
Joe Pesci Movies

Martin Scorsese Movies

Recommended to user

Our movie recommendation system employs a content-based approach, a sophisticated engine that employs additional user and movie information to curate personalized suggestions. At its core, the system utilizes cosine similarity, a fundamental mathematical concept, to quantify the likeness between user profiles and movie attributes.

Cosine Similarity and Personalization



Cosine similarity measures the angle between two vectors, reflecting their similarity. In our context, these vectors represent user profiles and movie attributes. By calculating the cosine of this angle, we gauge how closely a user's preferences align with specific movie features. Higher cosine similarity values indicate stronger alignment.

Precision and Insight

This approach not only predicts movies aligned with users' tastes but also explains the rationale behind these recommendations. It quantifies the extent of similarity, making the process transparent and interpretable. This fusion of prediction and explanation enhances the cinematic experience, offering tailored suggestions with clarity.

Enhancing User Engagement

Our content-based model, driven by cosine similarity, elevates user satisfaction by pinpointing movies that harmonize with individual preferences. Each recommendation becomes a guided journey, navigating the vast array of choices with finesse. Cosine similarity acts as our guiding compass, steering users toward cinematic gems that resonate deeply.

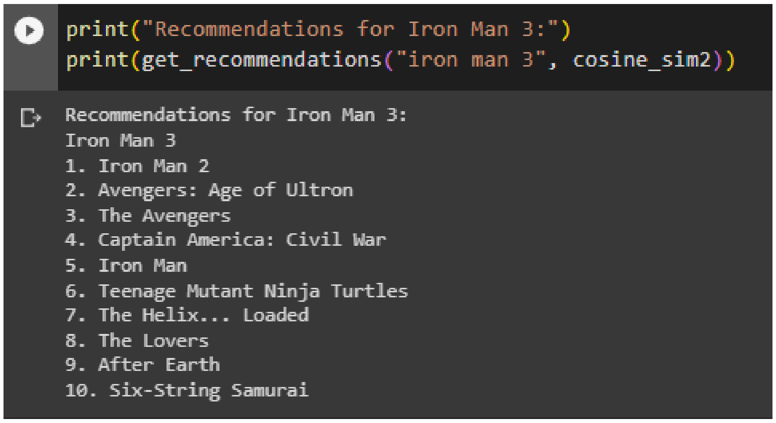
To wrap up this part, Incorporating cosine similarity into our content-based movie recommendation system enriches personalization and understanding. This method ensures precision, empowers users with insightful suggestions, and fosters a more engaging and fulfilling cinematic exploration.

Evaluation of Model Performance:

To assess the effectiveness of our movie recommendation model, we conducted a manual evaluation by scrutinizing the similarity between recommended movies and the user's preferred film. Specifically, we examined how closely the recommended movies align with the attributes of the user's designated favorite movie.

Case 1: Recommendations for "Iron Man 3"

Preferred Movie: "Iron Man 3"



```
print("Recommendations for Iron Man 3:")
print(get_recommendations("iron man 3", cosine_sim2))
```

```
Recommendations for Iron Man 3:
Iron Man 3
1. Iron Man 2
2. Avengers: Age of Ultron
3. The Avengers
4. Captain America: Civil War
5. Iron Man
6. Teenage Mutant Ninja Turtles
7. The Helix... Loaded
8. The Lovers
9. After Earth
10. Six-String Samurai
```

The movie "Iron Man 3" is renowned for its captivating blend of science fiction and action genres, characterized by keywords such as 'superheroes,' 'evil enemies,' and 'battles.' Our recommendation system admirably captures these thematic elements in its suggestions, thus reinforcing the system's efficacy.

1. "Iron Man 2": The sequel to the preferred movie, sharing the same protagonist and thematic essence, showcases our model's ability to discern continuity in narrative and genre.

2. "Avengers: Age of Ultron": As a pivotal installment in the Marvel Cinematic Universe, this recommendation resonates with the shared theme of superhero alliances and dynamic action sequences.

3. "The Avengers": A cornerstone of the superhero genre, this film offers a harmonious convergence of characters and action, in line with the user's cinematic inclination.

4. "Captain America: Civil War": Continuing the interconnected superhero saga, this suggestion underscores our model's astute recognition of genre continuity and thematic coherence.

5. "Iron Man": The origin story that laid the foundation for the user's preferred movie, offering a congruent blend of science fiction and action elements.

6. "Teenage Mutant Ninja Turtles": While not directly related to the "Iron Man" franchise, the recommendation acknowledges the shared affinity for action-packed narratives and dynamic characters.

7. "The Helix... Loaded": With a hint of science fiction and intrigue, this suggestion demonstrates the system's ability to recognize nuanced thematic parallels.

8. "The Lovers": While diverging in genre, this recommendation underscores the inherent challenge of fully aligning every attribute yet acknowledges the user's broader cinematic interests.

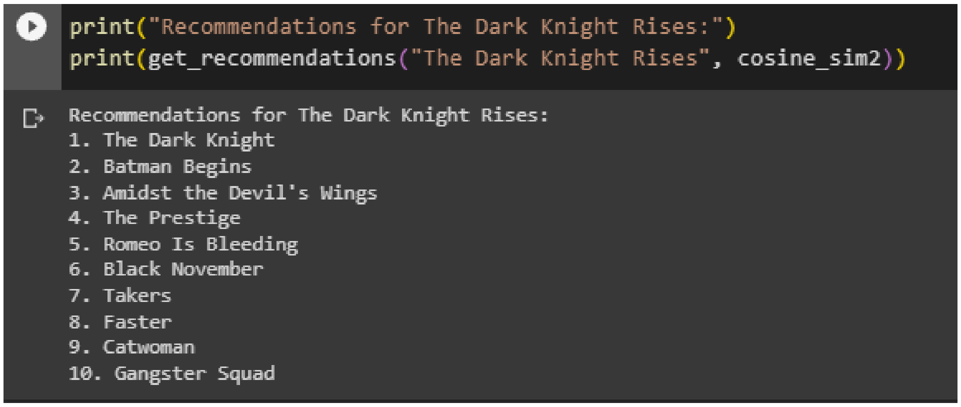
9. "After Earth": Exploring a science fiction landscape, this choice tangentially resonates with the user's genre preference.

10. "Six-String Samurai": By acknowledging the penchant for unique concepts and action elements, this suggestion extends the cinematic horizon.

In summation, our movie recommendation model astutely identifies and aligns with the genre, keywords, and thematic elements that define "Iron Man 3." While not every recommendation seamlessly mirrors every attribute, the model consistently showcases a profound understanding of the user's cinematic taste, substantiating the integral role of genre, keywords, and thematic elements in our recommendation algorithm. This evaluation underscores the system's efficacy in delivering tailored cinematic experiences and facilitating exploration within the realm of individual preferences.

Case 2: Recommendations for "The Dark Knight Rises":

Preferred Movie: "The Dark Knight Rises"



```
print("Recommendations for The Dark Knight Rises:")
print(get_recommendations("The Dark Knight Rises", cosine_sim2))
```

```
Recommendations for The Dark Knight Rises:
1. The Dark Knight
2. Batman Begins
3. Amidst the Devil's Wings
4. The Prestige
5. Romeo Is Bleeding
6. Black November
7. Takers
8. Faster
9. Catwoman
10. Gangster Squad
```

"The Dark Knight Rises" is emblematic of the action and thriller genres, marked by intense sequences and intriguing plot dynamics. Our recommendation system adeptly captures these attributes in its suggestions, reinforcing its proficiency in understanding thematic preferences.

1. "The Dark Knight": The predecessor to the preferred movie, renowned for its action-packed narrative and intricate character dynamics, demonstrating the model's ability to identify thematic continuity.

2. "Batman Begins": The genesis of the acclaimed Batman trilogy, resonating with the user's taste for superhero-driven narratives and suspenseful storytelling.

3. "Amidst the Devil's Wings": A selection that captures the tension and intensity present in the preferred movie, showcasing the system's appreciation for similar thriller elements.

4. "The Prestige": With its enthralling storyline and suspenseful atmosphere, this suggestion aligns with the user's inclination for intricate narratives and engaging plots.

5. "Romeo Is Bleeding": A film that merges elements of action and thriller, demonstrating the model's astuteness in identifying multi-faceted genre preferences.

6. "Black November": While distinct in plot, the recommendation acknowledges the shared affinity for suspense and gripping narratives.

7. "Takers": Aligning with the user's taste for high-stakes action and intrigue, this suggestion offers a congruent cinematic experience.

8. "Faster": Capitalizing on the preference for intense action, this choice enhances the spectrum of the user's cinematic interests.

9. "Catwoman": A direct nod to the superhero universe, acknowledging the broader thematic canvas while accounting for unique character-driven narratives.

10. "Gangster Squad": While differing in genre, this recommendation broadens the horizons by encompassing engaging plotlines and dynamic characters.

In summary, the recommendations for "The Dark Knight Rises" vividly mirror the action and thriller elements that define the user's preferred movie. The system adeptly identifies and aligns with thematic attributes, offering a diverse array of selections that encapsulate the user's affinity for intense narratives, dynamic characters, and suspenseful storytelling. This evaluation showcases the model's prowess in delivering tailored cinematic suggestions that resonate with individual preferences, ultimately elevating the user's exploration within the cinematic realm.

Innovativeness: Project Contributions and Distinctive Elements

Our project introduces innovative elements that distinguish it within the realm of movie recommendation systems:

Hybrid Methodology: Unlike conventional approaches, we employ a hybrid technique combining content-based classification and clustering. This unique blend widens the scope of movie recommendations, ensuring diversity and depth.

Chatbot Integration: We pioneer the integration of our recommendation system into a chatbot's dialogue flow. This real-time interaction showcases adaptability and introduces a dynamic dimension to content delivery.

Cosine Similarity Precision: Our incorporation of cosine similarity enhances personalization by precisely quantifying user-movie alignment. This elevates the accuracy of recommendations, leading to a more tailored experience.

Thematic Understanding: Leveraging clustering, our system grasps movie themes and user preferences on a nuanced level. This deep comprehension refines suggestions beyond surface attributes.

Interdisciplinary Fusion: Our project amalgamates AI, data analytics, and natural language processing, exemplifying a holistic approach that addresses complex entertainment dynamics.

User-Centric Interface: The chatbot-driven design emphasizes user experience, providing a seamless and engaging platform for exploring cinematic options.

Conclusion:

Our movie recommendation system employs advanced technologies and intricate algorithms to provide personalized suggestions, transforming content interactions. Through a content-based approach and cosine similarity, the system predicts movies aligned with individual preferences.

Manual evaluations confirmed the system's accuracy, consistently capturing genre, themes, and characters in recommended movies.

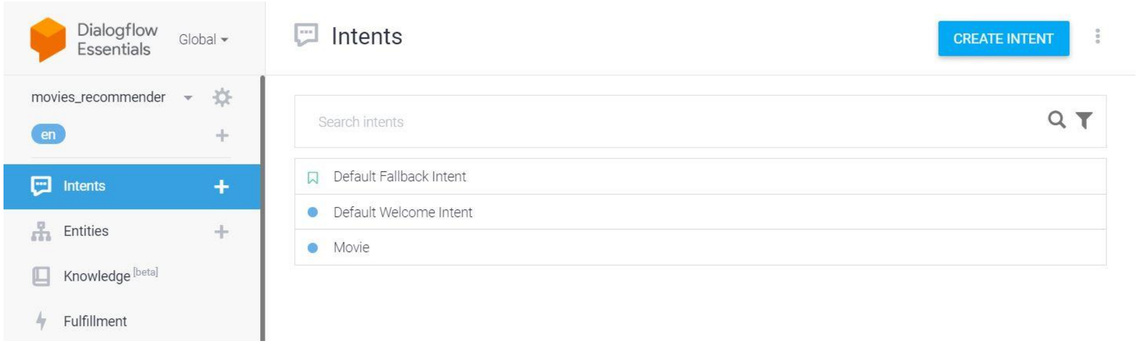
From dataset selection to algorithmic application, our foundation is strong. Guided by cosine similarity, the amalgamation of user attributes and movie features enhances engagement and insights.

Our system harmonizes film artistry and data science, guiding users to engaging movies. This project offers a glimpse into a future where tailored experiences reshape entertainment engagement.

Chatbot Implementation:

Chatbots have become essential tools in today's digital landscape. Their significance lies in their ability to provide instant assistance and engagement, enhancing customer experiences across industries. By automating routine tasks and inquiries, chatbots streamline processes, saving time and resources for businesses. They operate 24/7, ensuring round-the-clock availability, thereby increasing customer satisfaction and loyalty. Moreover, chatbots gather valuable data from interactions, enabling businesses to understand user preferences and tailor their services accordingly. They also offer scalability, handling multiple conversations simultaneously without compromising quality. In a world driven by convenience, chatbots stand as a pivotal bridge between businesses and customers, fostering efficient communication and problem-solving.

After we did the recommendation system and make sure it worked well, we started creating our chatbot. We started creating it using Dialogflow. We began to imagine possible scenarios and put them into consideration, as shown in the next figures:



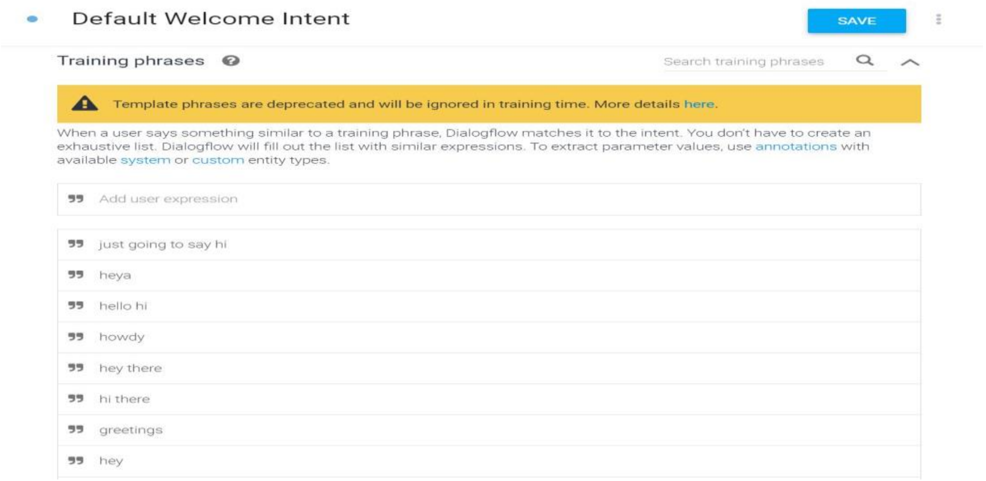
Default Fallback Intent

Default Welcome Intent

Movie

Our agent has three intents, first is “Default Fallback Intent” This one by default exists, The benefit of it is that when the user writes a message and the chatbot fails to respond to it, it shows the user a response through it, which are responses that express his lack of understanding of what the user said and asks him to put a clearer message.

The second is “Default Welcome Intent”, which is responsible for greeting the user, there is a set of messages that people start with the greeting, as well as the answers that the chatbot will respond to in order to complete the conversation with him, as shown in the following two figures:



Template phrases are deprecated and will be ignored in training time. More details here.

When a user says something similar to a training phrase, Dialogflow matches it to the intent. You don't have to create an exhaustive list. Dialogflow will fill out the list with similar expressions. To extract parameter values, use annotations with available system or custom entity types.

Add user expression

just going to say hi

heya

hello hi

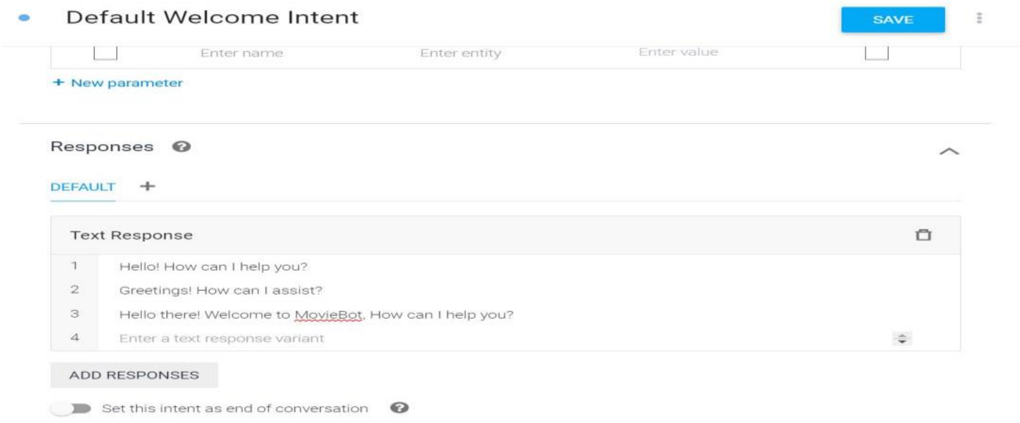
howdy

hey there

hi there

greetings

hey



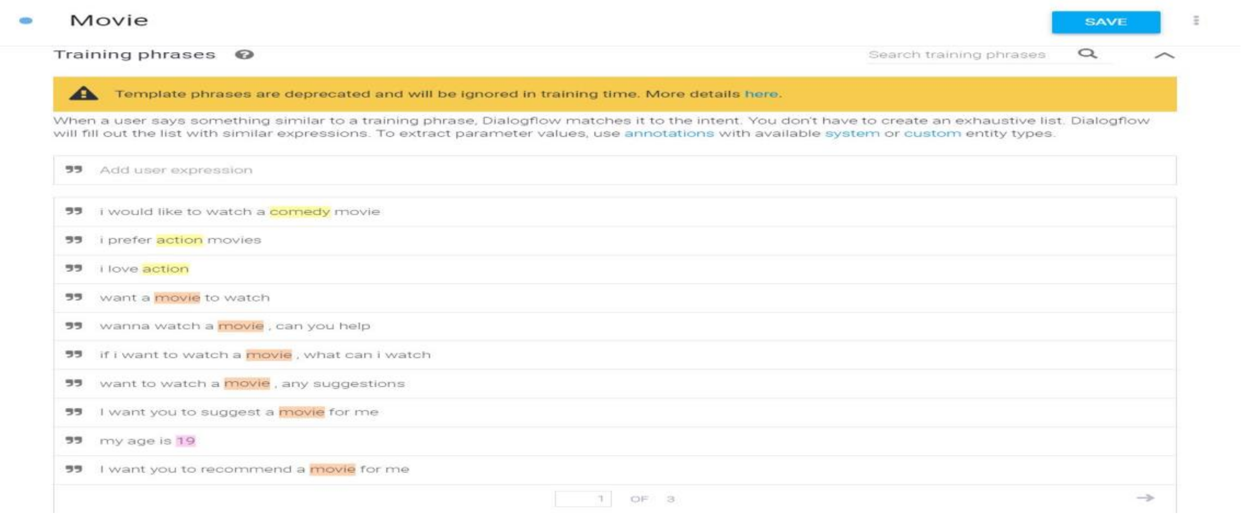
1 Hello! How can I help you?

2 Greetings! How can I assist?

3 Hello there! Welcome to MovieBot, How can I help you?

4 Enter a text response variant

We now come to the last Intent "Movie", Here we have put a set of questions and more than one way for the same answer. We did this based on our need for some information that we need to suggest the appropriate movie to the user.



Template phrases are deprecated and will be ignored in training time. More details here.

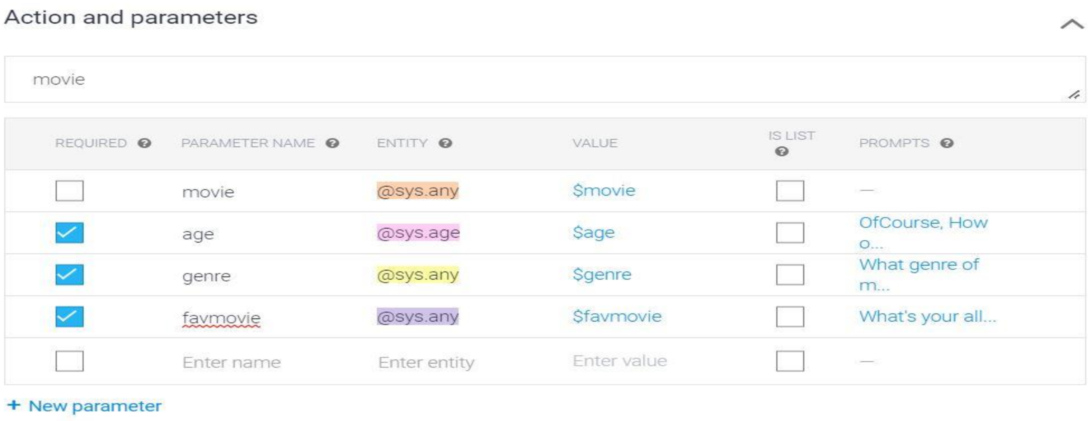
When a user says something similar to a training phrase, Dialogflow matches it to the intent. You don't have to create an exhaustive list. Dialogflow will fill out the list with similar expressions. To extract parameter values, use annotations with available system or custom entity types.

Add user expression

- i would like to watch a comedy movie
- i prefer action movies
- i love action
- want a movie to watch
- wanna watch a movie , can you help
- if i want to watch a movie , what can i watch
- want to watch a movie , any suggestions
- I want you to suggest a movie for me
- my age is 19
- I want you to recommend a movie for me

For example, in this figure, when the chatbot asked about the favorite genre, the user can respond in several ways as you see.

And we want to extract the important information from his response, so we need the parameters, parameters Play an important role because these are what we will send to our recommendation system, and based on it The appropriate movies will come in response.

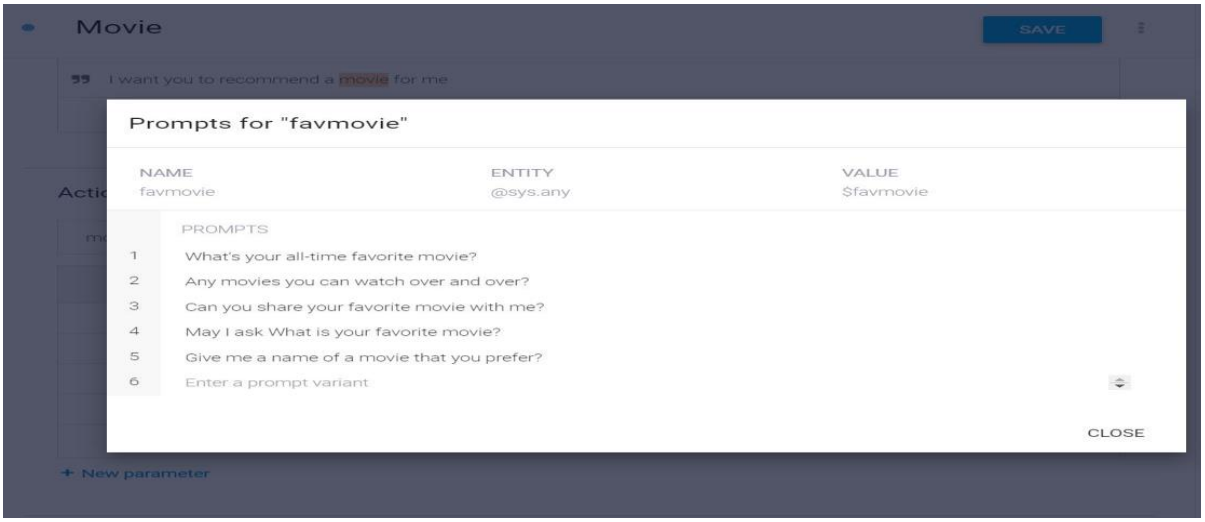


Action and parameters

movie

REQUIRED	PARAMETER NAME	ENTITY	VALUE	IS LIST	PROMPTS
□	movie	@sys.any	$movie	□	—
☑	age	@sys.age	$age	□	OfCourse, How o...
☑	genre	@sys.any	$genre	□	What genre of m...
☑	favmovie	@sys.any	$favmovie	□	What's your all...
□	Enter name	Enter entity	Enter value	□	—

As you see, These required parameters we need, and each one of them has prompts questions, that's the question the chatbot will ask the user and wait for the response.



" I want you to recommend a movie for me

Prompts for "favmovie"

NAME	ENTITY	VALUE
favmovie	@sys.any	$favmovie

PROMPTS

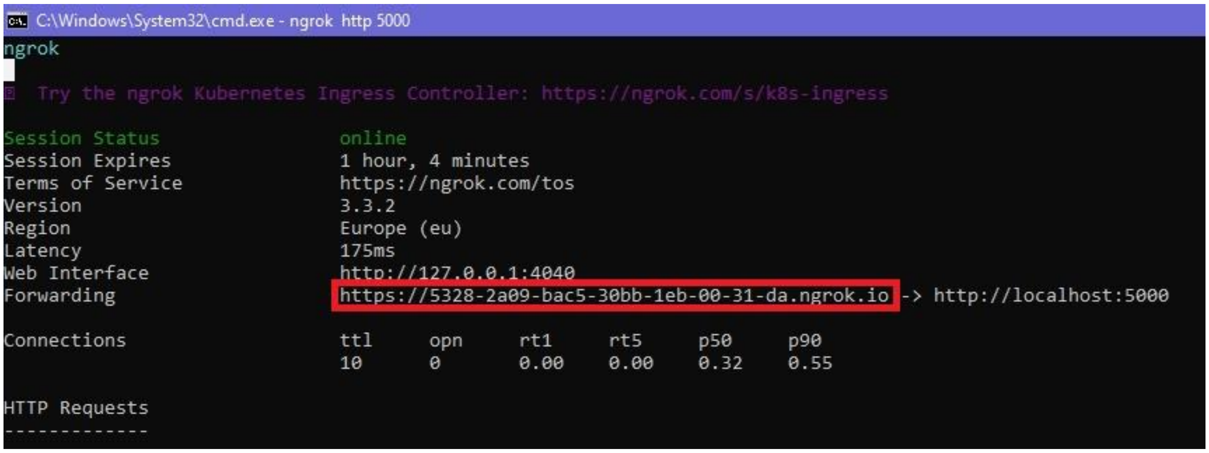
1 What's your all-time favorite movie?
2 Any movies you can watch over and over?
3 Can you share your favorite movie with me?
4 May I ask What is your favorite movie?
5 Give me a name of a movie that you prefer?
6 Enter a prompt variant

CLOSE

For example, here are some questions asked about favorite movies.

Now we want to respond to the user from our recommendation system, to do that we need to connect the chatbot with the recommendation system, by using Flask and ngrok.

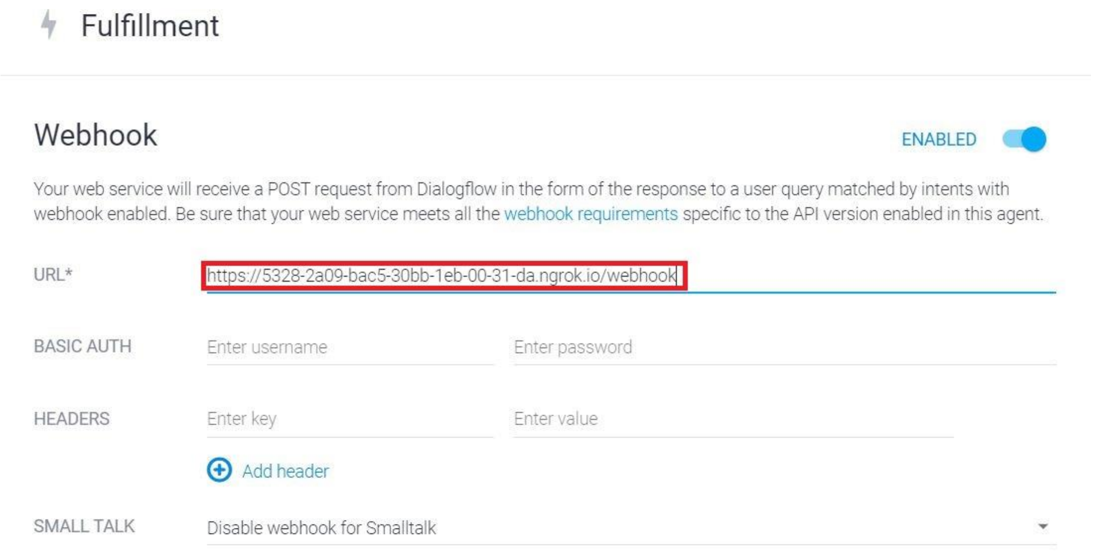
We need to run local server by using ngrok, via this command (ngrok http 5000) after that this is the result:



C:\Windows\System32\cmd.exe - ngrok http 5000

```
ngrok

   Try the ngrok Kubernetes Ingress Controller: https://ngrok.com/s/k8s-ingress

Session Status                online
Session Expires               1 hour, 4 minutes
Terms of Service              https://ngrok.com/tos
Version                       3.3.2
Region                        Europe (eu)
Latency                       175ms
Web Interface                 http://127.0.0.1:4040
Forwarding                    https://5328-2a09-bac5-30bb-1eb-00-31-da.ngrok.io -> http://localhost:5000

Connections                   ttl     opn     rt1     rt5     p50     p90
                              10      0       0.00    0.00    0.32    0.55

HTTP Requests
-------------
```

Then we need this url, and put it in dialogflow in fulfilment section like this:



Fulfillment

Webhook

ENABLED

Your web service will receive a POST request from Dialogflow in the form of the response to a user query matched by intents with webhook enabled. Be sure that your web service meets all the webhook requirements specific to the API version enabled in this agent.

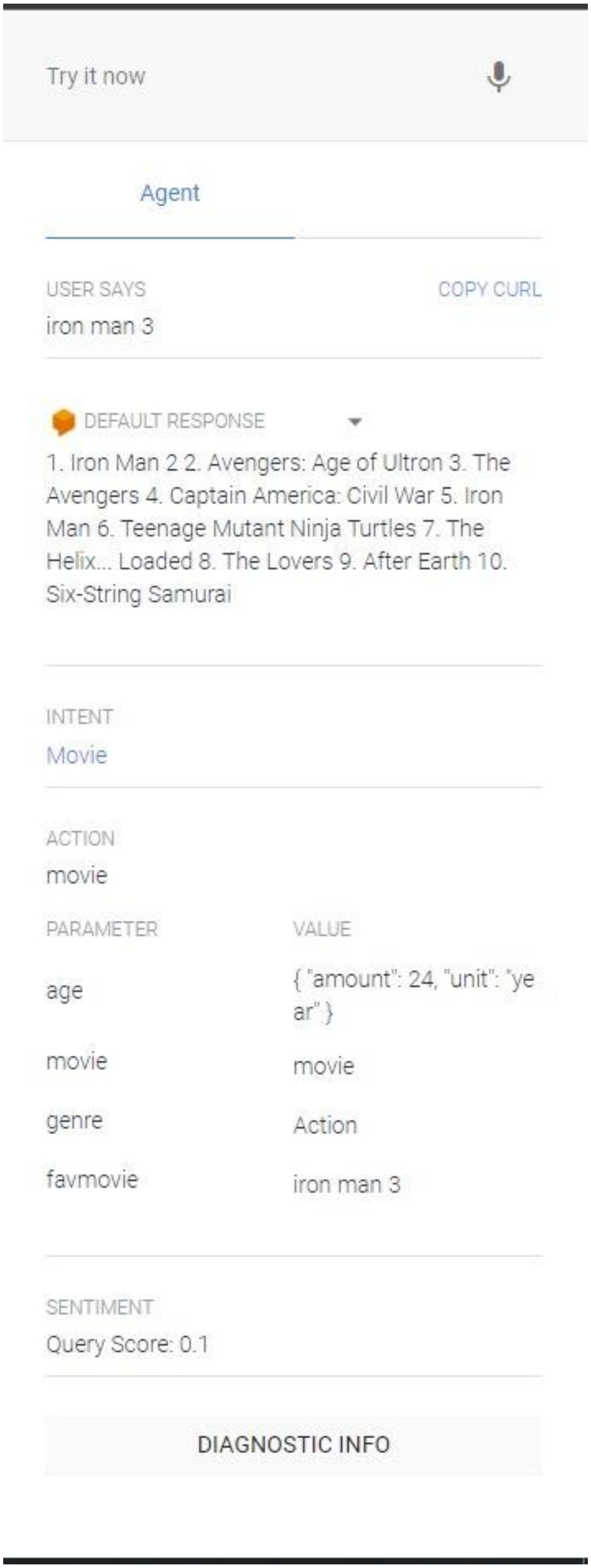
URL* https://5328-2a09-bac5-30bb-1eb-00-31-da.ngrok.io/webhook

BASIC AUTH Enter username Enter password

HEADERS Enter key Enter value

SMALL TALK Disable webhook for Smalltalk

Results:



USER SAYS
iron man 3

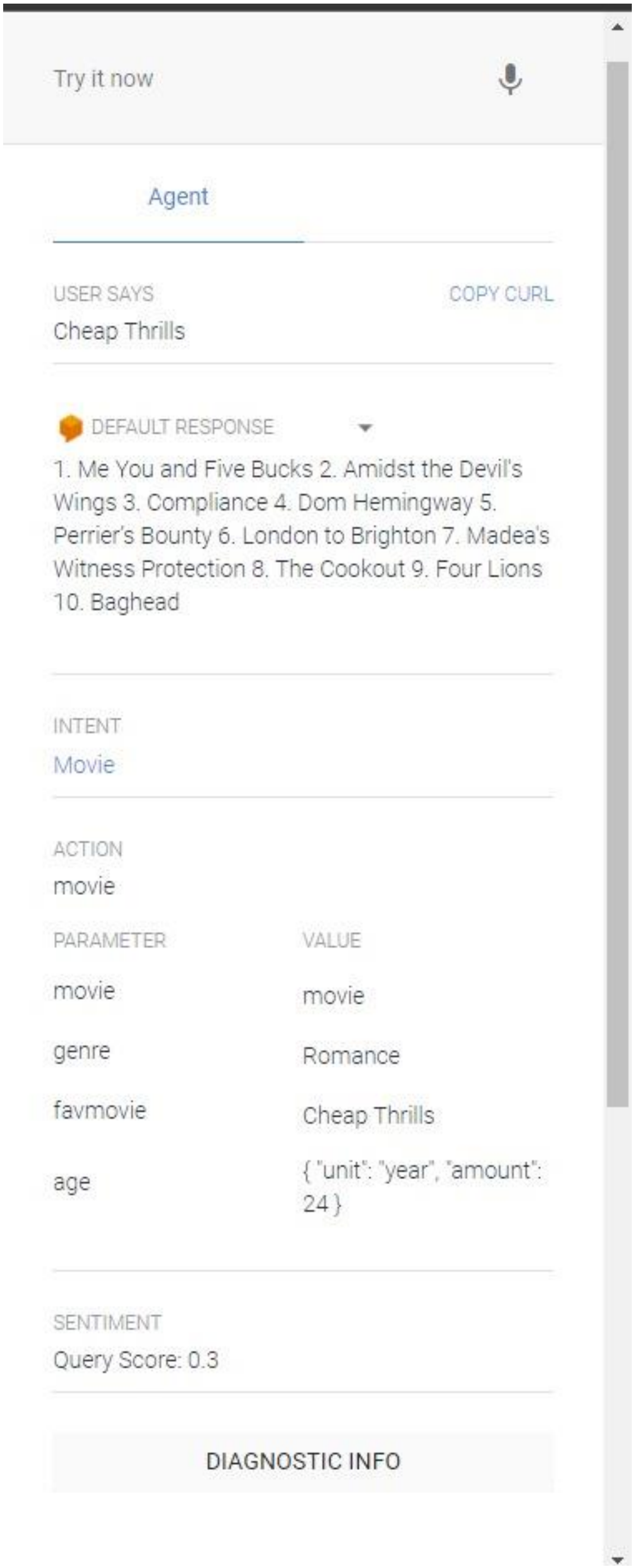
DEFAULT RESPONSE
1. Iron Man 2 2. Avengers: Age of Ultron 3. The Avengers 4. Captain America: Civil War 5. Iron Man 6. Teenage Mutant Ninja Turtles 7. The Helix... Loaded 8. The Lovers 9. After Earth 10. Six-String Samurai

INTENT
Movie

ACTION
movie

PARAMETER	VALUE
age	{ "amount": 24, "unit": "year" }
movie	movie
genre	Action
favmovie	iron man 3

SENTIMENT
Query Score: 0.1



USER SAYS
Cheap Thrills

DEFAULT RESPONSE
1. Me You and Five Bucks 2. Amidst the Devil's Wings 3. Compliance 4. Dom Hemingway 5. Perrier's Bounty 6. London to Brighton 7. Madea's Witness Protection 8. The Cookout 9. Four Lions 10. Baghead

INTENT
Movie

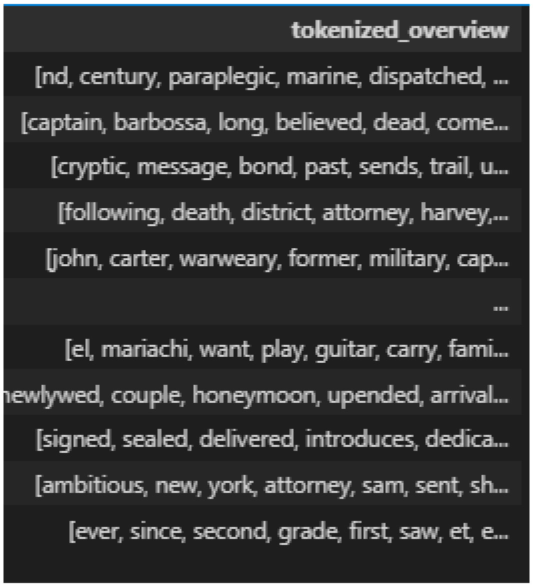
ACTION
movie

PARAMETER	VALUE
movie	movie
genre	Romance
favmovie	Cheap Thrills
age	{ "unit": "year", "amount": 24 }

SENTIMENT
Query Score: 0.3

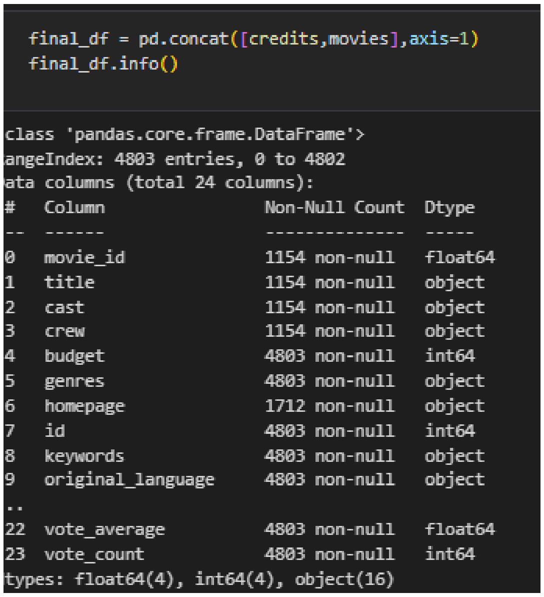
Classification and Clustering:

Data Analysis and cleaning:



```
                          tokenized_overview
 [nd, century, paraplegic, marine, dispatched, ...
  [captain, barbossa, long, believed, dead, come...
     [cryptic, message, bond, past, sends, trail, u...
       [following, death, district, attorney, harvey,...
       [john, carter, warweary, former, military, cap...
                                                   ...
          [el, mariachi, want, play, guitar, carry, fami...
ewlywed, couple, honeymoon, upended, arrival...
 [signed, sealed, delivered, introduces, dedica...
 [ambitious, new, york, attorney, sam, sent, sh...
      [ever, since, second, grade, first, saw, et, e...
```

Figure 2Data post-cleaning and NLP processing



```
  final_df = pd.concat([credits,movies],axis=1)
  final_df.info()

class 'pandas.core.frame.DataFrame'>
angeIndex: 4803 entries, 0 to 4802
ata columns (total 24 columns):
#   Column             Non-Null Count  Dtype
--  ------             --------------  -----
0   movie_id           1154 non-null   float64
1   title              1154 non-null   object
2   cast               1154 non-null   object
3   crew               1154 non-null   object
4   budget             4803 non-null   int64
5   genres             4803 non-null   object
6   homepage           1712 non-null   object
7   id                 4803 non-null   int64
8   keywords           4803 non-null   object
9   original_language  4803 non-null   object
..
22  vote_average       4803 non-null   float64
23  vote_count         4803 non-null   int64
types: float64(4), int64(4), object(16)
```

Figure 1Some info about data

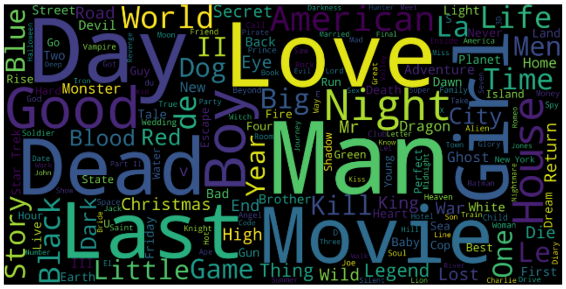


Figure 3Most words in overview

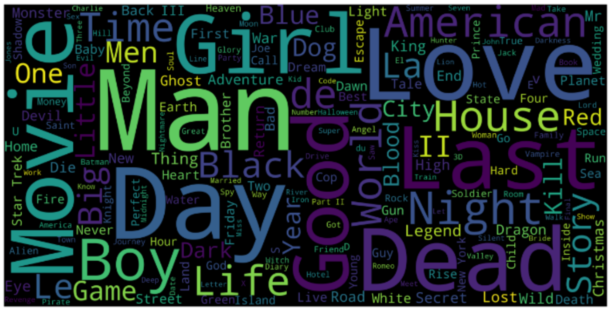


Figure 4Most words in keywords

Feature Engineering

We used TF-IDF and Bag-of-Words (BoW) techniques to represent the data in a numerical format. After that, we employed T-SNE visualization to gain insights into the data's complexity and identify potential models for further analysis. This process allowed us to understand the underlying patterns and structures in the data, making it easier to select appropriate models and algorithms for clustering and classification tasks. The combination of these methods proved to be effective in enhancing data understanding and facilitating the extraction of valuable information from the dataset.

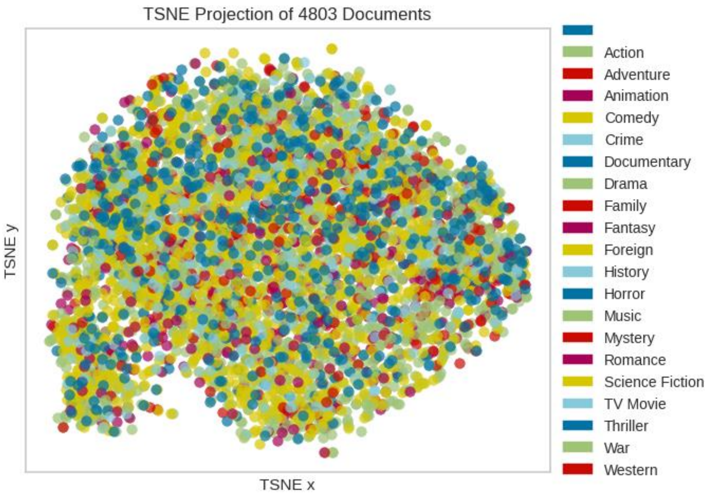


Figure 5plot the T-sne distribution for the bow Dataframe

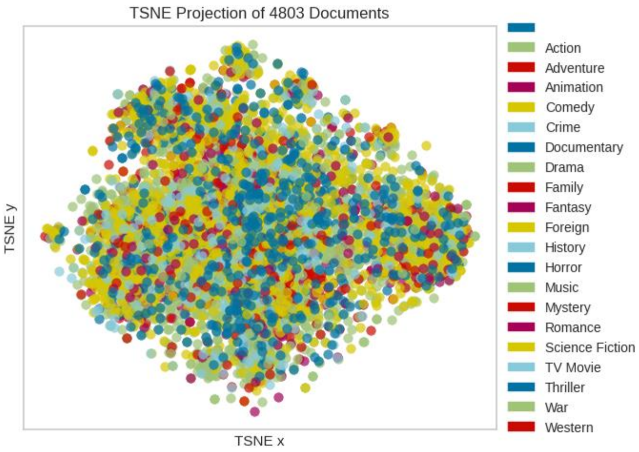
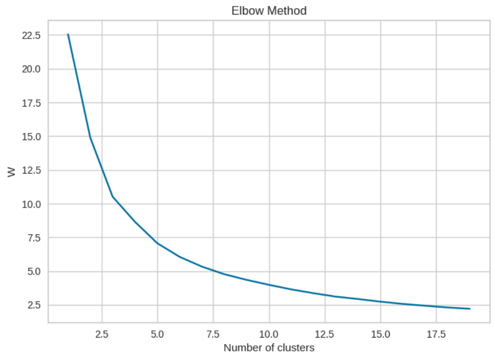


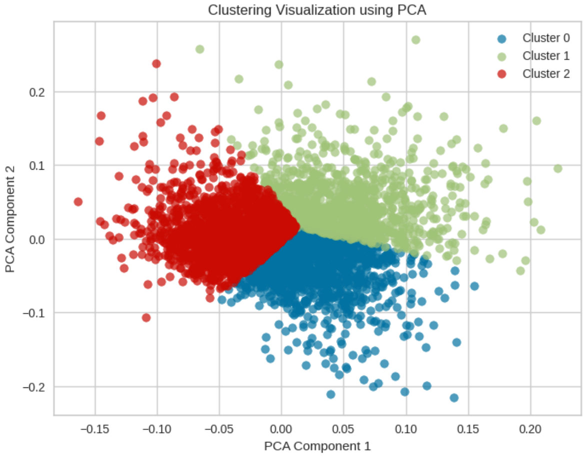
Figure 6plot the T-sne distribution for the TFIDF Dataframe

Clustering

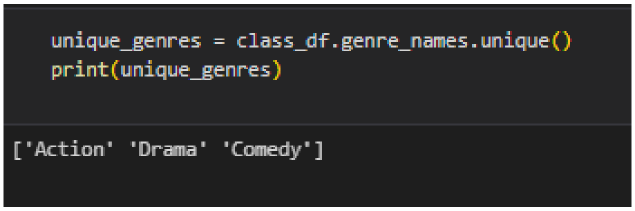
After completing the initial feature engineering and data preprocessing steps, we sought to gain a deeper understanding of the underlying patterns within the dataset. To achieve this, we applied clustering techniques, specifically the k-means algorithm. This powerful tool in unsupervised learning allowed us to identify and group similar samples together, revealing hidden structures and patterns within the data. By doing so, we gained valuable insights into the dataset's inherent relationships and distinct clusters, which can aid in further analysis and decision-making processes.



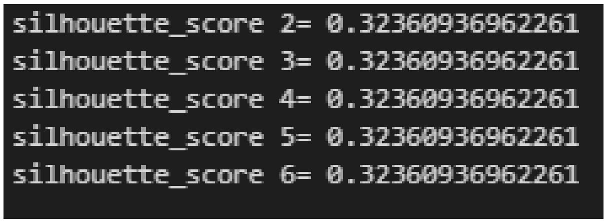
Using the elbow method, we determined the optimal number of clusters that best captured the inherent patterns in the data. This technique involved plotting the variance explained by different numbers of clusters and selecting the point at which adding more clusters provided diminishing returns in terms of variance explained. As a result, we discovered that the data naturally formed three distinct clusters.



We have identified three separate genre clusters in the dataset.



```
    unique_genres = class_df.genre_names.unique()
    print(unique_genres)

['Action' 'Drama' 'Comedy']
```



```
silhouette_score 2= 0.32360936962261
silhouette_score 3= 0.32360936962261
silhouette_score 4= 0.32360936962261
silhouette_score 5= 0.32360936962261
silhouette_score 6= 0.32360936962261
```

Figure 3This is an evaluation of the k-means clustering.

This clustering information proved to be invaluable in enhancing our subsequent classification models. By having a clearer view of the data's inherent groupings, we could fine-tune our feature representations and model architectures accordingly. This, in turn, led to improved performance and more accurate predictions in our classification tasks.

By combining clustering with our existing feature engineering efforts, we were able to extract valuable insights and gain a comprehensive understanding of the data's underlying structure. This iterative approach not only strengthened our models but also allowed us to uncover potential relationships and correlations among the data points, ultimately empowering us to make more informed decisions in our analysis and predictions.

Classification

After identifying the clusters through the previously mentioned process, we proceeded to select several classification models, namely SVM, Random Forest, Decision Tree, and Logistic Regression, for training.

Each of these models has its strengths and characteristics, which allowed us to gain diverse insights into the data.

To assess the performance of these models, we utilized confusion matrices.

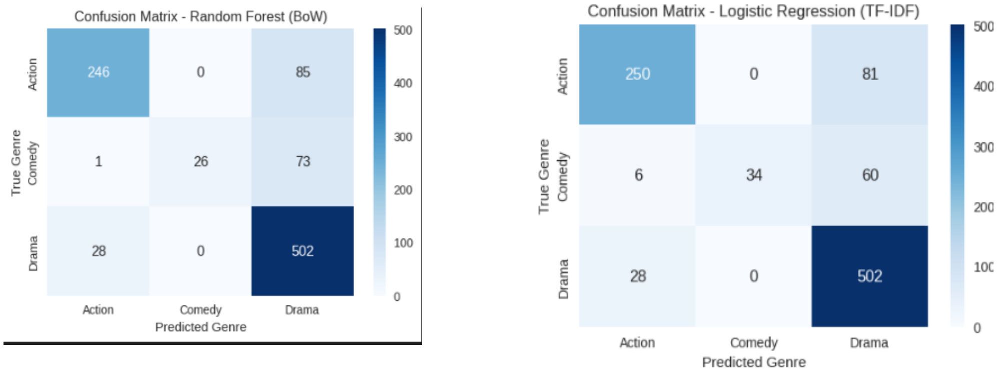


Figure 7Confusion Matrix for Logistic Regression (TF-IDF) & (BOW)

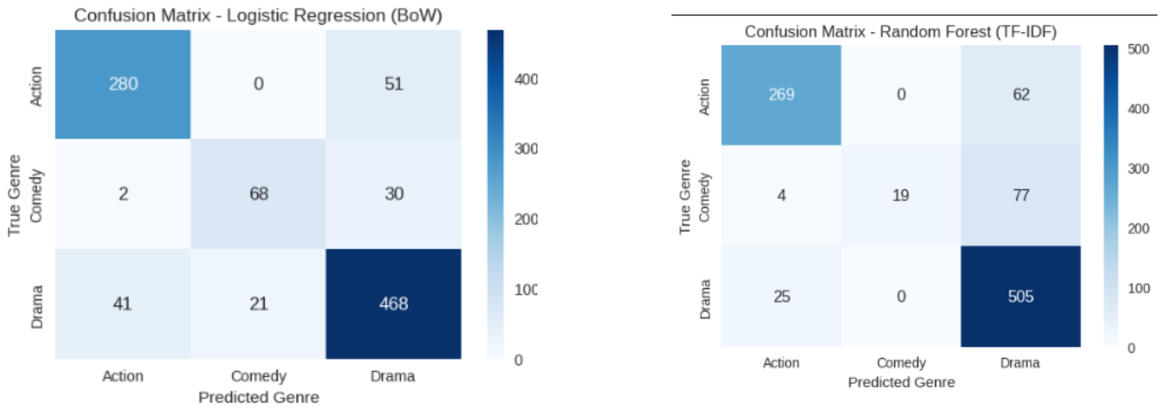


Figure 8Confusion Matrix for Random Forest (TF-IDF) & (BOW)

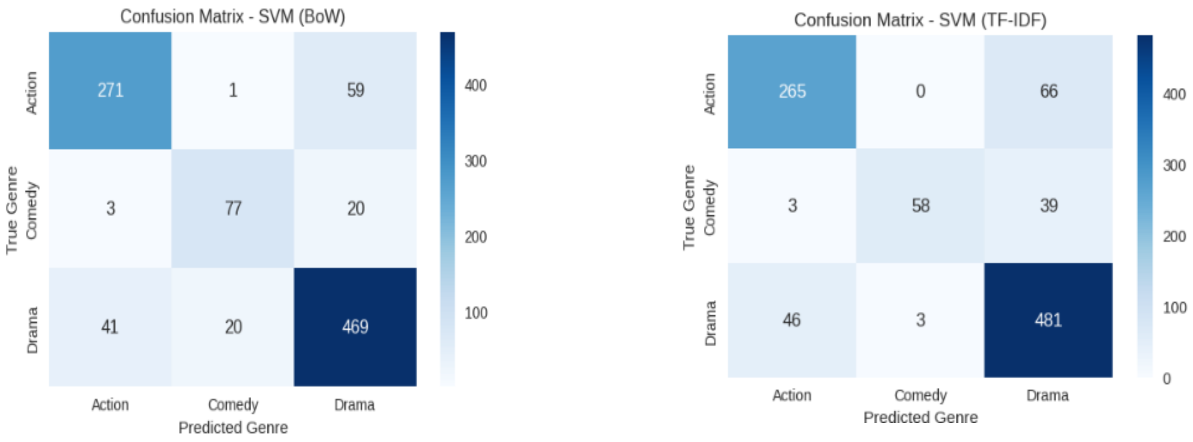


Figure 9Confusion Matrix for SVM (TF-IDF) & (BOW)

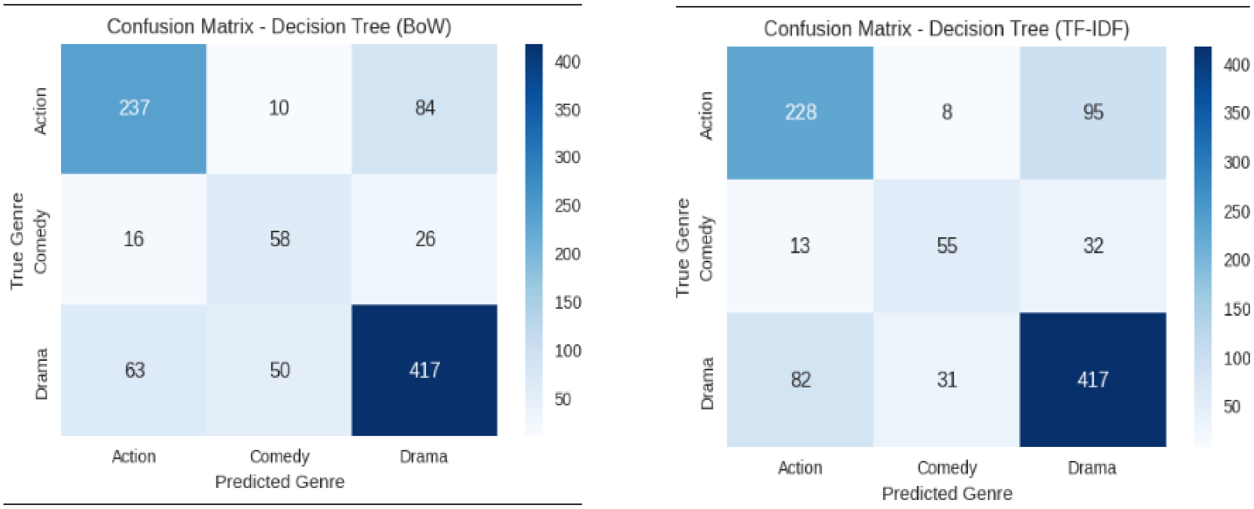
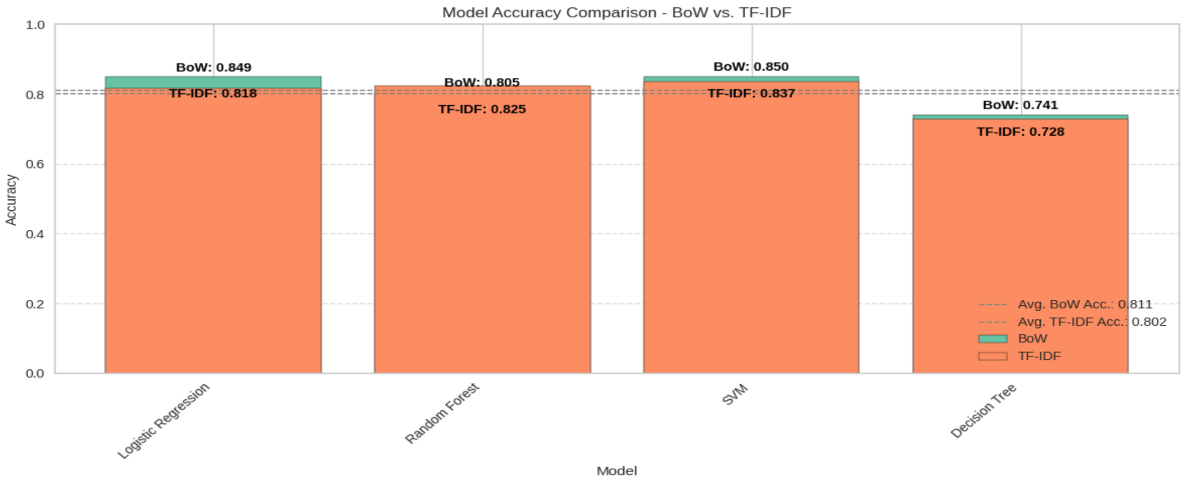
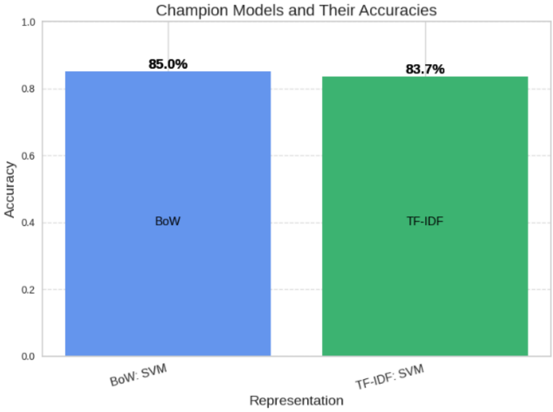


Figure 10Confusion Matrix for SVM (TF-IDF) & (BOW)

Moreover, we performed a visual comparison of the accuracies obtained from the TF-IDF and BoW models. This comparison allowed us to analyze their relative performance and determine which data representation method was more effective in our specific context.



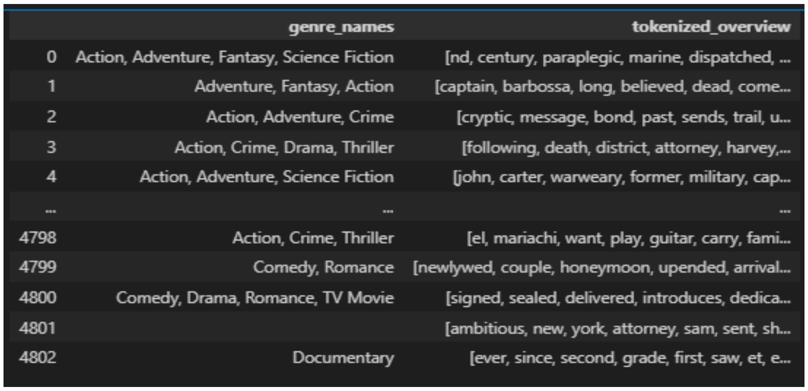
Following a thorough evaluation and comparison, we arrived at a clear conclusion. The SVM model emerged as the best-performing classifier among all the models tested. With its highest accuracy score, as indicated in the chart, SVM demonstrated its ability to effectively classify movie overviews into the correct genres.



this comprehensive evaluation process, involving multiple models, confusion matrices, and visual comparisons, enabled us to confidently select SVM as the optimal model for our specific task, ensuring accurate genre predictions and enhancing the overall quality of our analysis.

Error analysis

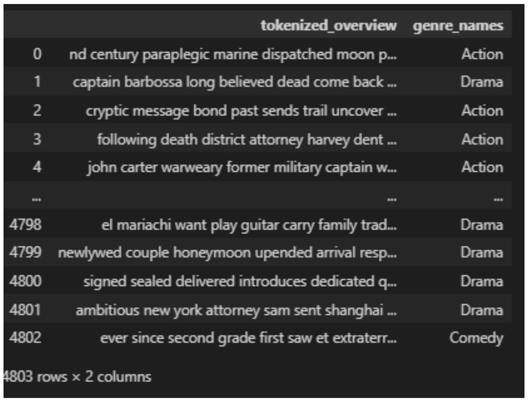
The error analysis for classification revealed that the weak accuracy of the models was primarily due to the challenge posed by movie overviews being associated with multiple genres. This genre overlap introduced ambiguity and made it challenging for the classification models to make precise predictions.



	genre_names	tokenized_overview
0	Action, Adventure, Fantasy, Science Fiction	[nd, century, paraplegic, marine, dispatched, ...
1	Adventure, Fantasy, Action	[captain, barbossa, long, believed, dead, come...
2	Action, Adventure, Crime	[cryptic, message, bond, past, sends, trail, u...
3	Action, Crime, Drama, Thriller	[following, death, district, attorney, harvey,...
4	Action, Adventure, Science Fiction	[john, carter, warweary, former, military, cap...
...	...	...
4798	Action, Crime, Thriller	[el, mariachi, want, play, guitar, carry, fami...
4799	Comedy, Romance	[newlywed, couple, honeymoon, upended, arrival...
4800	Comedy, Drama, Romance, TV Movie	[signed, sealed, delivered, introduces, dedica...
4801		[ambitious, new, york, attorney, sam, sent, sh...
4802	Documentary	[ever, since, second, grade, first, saw, et, e...

To tackle this complexity, we used clustering techniques to group similar movie overviews together. This allowed us to find natural patterns and themes within the data.

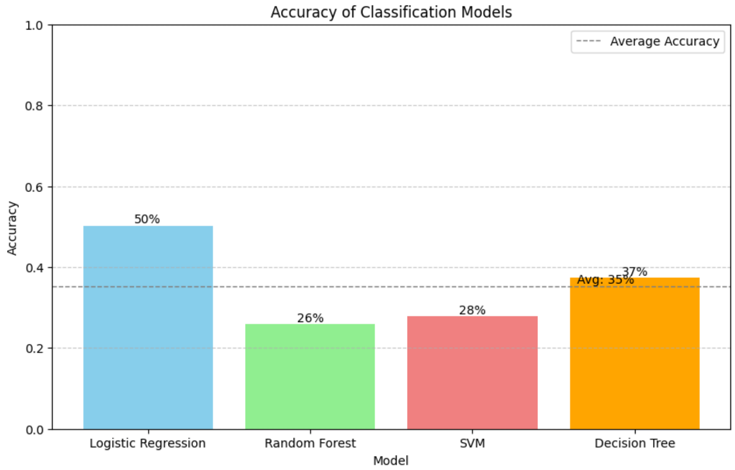
After clustering, we focused on the most common genre in each cluster and adjusted the data accordingly. This helped us simplify the information and provide clearer input to the classification models



	tokenized_overview	genre_names
0	nd century paraplegic marine dispatched moon p...	Action
1	captain barbossa long believed dead come back ...	Drama
2	cryptic message bond past sends trail uncover ...	Action
3	following death district attorney harvey dent ...	Action
4	john carter warweary former military captain w...	Action
...	...	...
4798	el mariachi want play guitar carry family trad...	Drama
4799	newlywed couple honeymoon upended arrival resp...	Drama
4800	signed sealed delivered introduces dedicated q...	Drama
4801	ambitious new york attorney sam sent shanghai ...	Drama
4802	ever since second grade first saw et extraterr...	Comedy

4803 rows × 2 columns

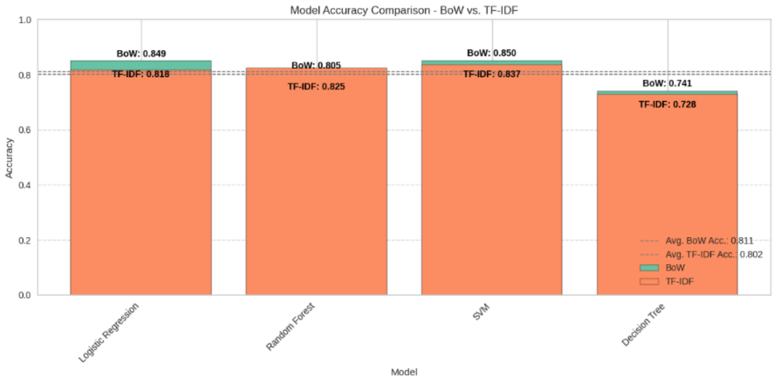
Based on the chart provided, we noticed that the accuracy was low prior to implementing this technique.



By refining the data and associating each cluster with its dominant genre, we reduced the confusion caused by genre overlap. This, in turn, improved the models' ability to accurately predict genres for new movie overviews.

Through this approach of clustering and data modification, we gained valuable insights into the data and resolved the genre overlap challenge. The result was more reliable and accurate classification models, ensuring better predictions for any movie overview in the future.

After implementing this technique, we observed a significant increase in accuracy, as indicated by the chart.


Model Accuracy Comparison - BoW vs. TF-IDF
Accuracy
BoW: 0.849
TF-IDF: 0.818
BoW: 0.805
TF-IDF: 0.825
BoW: 0.850
TF-IDF: 0.837
BoW: 0.741
TF-IDF: 0.728
Avg. BoW Acc.: 0.811
Avg. TF-IDF Acc.: 0.802
BoW
TF-IDF
Logistic Regression
Random Forest
SVM
Decision Tree
Model