Natural language Processing Group Project Final Report

Title: Netflix Recommendation System & Chatbot

Team 8

Team Members:

Eun Hye Kim, Made Jessica Felicia, Yi-En Liu, Hsing-Chia Tsai, Chieh-Yi Chou

Table of Contents

1. Introduction

This report presents the clustering analysis conducted for a personalized recommendation system for Netflix movies and TV shows. The goal is to group similar content based on their descriptions, genres, and other metadata using Natural Language Processing (NLP) techniques. In addition to developing a recommendation system, this project also includes the implementation of a chatbot that can interact with users, providing personalized movie and TV show suggestions based on their preferences. The chatbot leverages NLP to understand user queries, analyze content similarity, and deliver relevant recommendations in a conversational format.

2. Project Objectives

- Develop a content-based recommendation system: Utilize machine learning and NLP techniques to generate personalized recommendations based on user preferences and content similarity.
- Implement NLP techniques for text analysis: Extract meaningful insights from descriptions, genres, and cast information using text preprocessing and vectorization techniques.
- Enhance user experience with a chatbot: Integrate a chatbot that can provide movie and TV show suggestions based on conversational inputs, making the recommendation process more interactive.
- Optimize clustering methods for content categorization: Use unsupervised learning techniques such as K-Means clustering to group similar content, improving the accuracy of recommendations.
- Evaluate model effectiveness and user engagement: Assess the performance of the recommendation system and chatbot through qualitative and quantitative measures, ensuring an optimal user experience.

3. Data Pre-processing

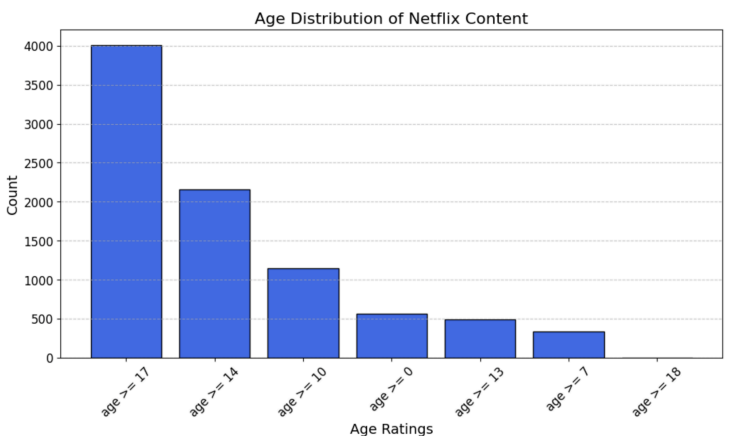
3.1 Missing Value Analysis

As part of the data preprocessing step, the dataset was examined for missing values. A column-wise check was performed to identify any features with incomplete data. The analysis revealed that the dataset contained no missing values across all features. This result indicates that the dataset is complete and requires no removal of rows or columns due to missing entries. The completeness of the dataset ensures reliability for further analysis.

cleaned words in a new column, "tokenized_description". This transformation refines raw text into structured data, making it more suitable for further analysis like trend identification, sentiment classification, or recommendation systems.

index	title	tokenized_description
0	Dick Johnson Is Dead	father,nears,end,life,filmmaker,kirsten,johnson,stage,death,inventive,comical,way,help,face,inevitable
1	Blood & Water	crossing,path,party,cape,town,teen,set,prove,whether,swimming,star,sister,abducted,birth
2	Ganglands	protect,family,powerful,drug,lord,skilled,thief,mehdi,expert,team,robber,pulled,violent,deadly,turf,war
3	Jailbirds New Orleans	feud,flirtation,toilet,talk,go,among,incarcerated,woman,orleans,justice,center,new,orleans,gritty,reality,series
4	Kota Factory	city,coaching,center,known,train,finest,collegiate,mind,earnest,unexceptional,student,friend,navigate,campus,life

3.2 Age Rating Classification



The age_ratings_numeric dictionary maps movie and TV content ratings to their respective age restrictions, defining the minimum recommended viewing age for each category. Movie ratings include "G" and "TV-Y" for all ages, "PG" and "TV-PG" for audiences 10 years and older, "PG-13" for 13 and older, "R" and "TV-MA" for 17 and older, and "NC-17" for adults 18 and above. TV-specific ratings such as "TV-Y7" and "TV-Y7-FV" set the minimum age at 7, while "TV-14" restricts content to viewers 14 and

older. This mapping facilitates structured age-based filtering, ensuring appropriate content viewing based on age restrictions.

3.3 Data Tokenization

We performed tokenization to extract meaningful words from the "description" column, aiming to prepare the text for further analysis such as sentiment detection or categorization. To achieve this, we used NLTK's text processing tools, ensuring a structured and clean dataset. Our approach involved several key steps: first, we converted text to lowercase for uniformity, then we tokenized it into individual words. To remove unnecessary noise, we filtered out punctuation, non-alphanumeric characters, and stopwords—common words that don't contribute much meaning. Finally, we lemmatized the words, reducing them to their base forms to ensure consistency. We then applied this process to every row in the "description" column, storing the

Utilizes TF-IDF vectorization to analyze text descriptions and extract key terms. It applies TfidfVectorizer to transform the text into a (8798, 15314) matrix. Then, it calculates the average TF-IDF weight of words and selects the 20 most representative terms, such as "life," "young," "family," "woman", highlighting words that are most distinctive for text classification. This process helps machine learning models understand the core themes of the text and improve analysis accuracy.

4. Model Description

We employed three different machine learning models to predict the genre of movies or TV shows, achieving the following F1-micro average accuracy scores: Logistic Regression + SVD (Singular Value Decomposition) at 0.3544 ± 0.0153, Multinomial Naive Bayes + NMF (Non-Negative Matrix Factorization) at 0.1816 ± 0.0116, and Random Forest + SVD (Singular Value Decomposition) at 0.3382 ± 0.0128. The overall lower accuracy is primarily due to the fact that our predictions were based on the tokenized description field, while both description and genre contain multiple values in our dataset. This made it challenging for the models to accurately perform multi-label classification, as they struggled to learn the relationships between different genres. Additionally, many movie and TV descriptions are broad and applicable to multiple genres, making it difficult for the models to distinguish between categories, ultimately impacting predictive performance.

```
Model: LogisticRegression+SVD
F1-micro (3-fold avg): 0.3544 ± 0.0153
------------------------------------------------
Model: MultinomialNB+NMF
F1-micro (3-fold avg): 0.1816 ± 0.0116
------------------------------------------------
Model: RandomForest+SVD
F1-micro (3-fold avg): 0.3382 ± 0.0128
------------------------------------------------
```

5. Methodology

The process of developing the recommendation system involved several key steps, including data organization, content categorization, and interactive search optimization. To identify patterns within the dataset, a method was employed to group similar titles based on their content. The most appropriate number of groups was determined through comparative evaluation, ensuring that each category represented a distinct type of content. The defining characteristics of each group were then examined by analyzing common terms and themes, which provided insight into the nature of the categorized titles. An interactive component was introduced to improve user engagement with the system. This feature processes user inputs, identifies key preferences such as genre or specific themes, and retrieves suggestions based on the established content groups. By combining structured categorization with an intuitive search function, the system offers a streamlined way for users to find relevant titles. The overall approach ensures that recommendations are tailored to individual interests while maintaining an efficient and user-friendly experience. To make the recommendation process more interactive, a chatbot was integrated as the user-facing interface. The chatbot processes natural language inputs, guiding users through structured queries and dynamically adjusting recommendations based on conversational interactions. It allows users to discover new content through an engaging and intuitive experience rather than relying solely on static search filters. By combining machine learning, clustering, and interactive query processing, the system delivers a streamlined and personalized approach to Netflix content discovery.

6. Clustering Results

Category 1: Big City Life & Personal Discovery

This category includes stories set in bustling urban environments, often exploring themes of ambition, self-growth, and navigating social dynamics. These narratives frequently follow characters experiencing major life changes, adjusting to new environments, or striving for success in their personal or professional lives.

- Examples: Blood & Water, Ganglands, Jailbirds New Orleans
- Common Themes: Urban life, self-discovery, career aspirations, social challenges

Category 2: Love & Life Journeys

Titles in this category revolve around relationships, romance, and personal evolution. These stories explore the emotional highs and lows of love, friendships, and the transformative power of human connection.

- Examples: Midnight Mass, The Starling, Je Suis Karl
- Common Themes: Romance, relationships, heartbreak, emotional growth, life-changing events

Category 3: Coming-of-Age & Life Transitions

This category consists of films and shows that depict personal growth, adolescence, and the transition from one phase of life to another. The characters often face challenges related to identity, education, or societal expectations, making these narratives highly relatable.

- Examples: Kota Factory, Bright Star, Dhanak
- Common Themes: Adolescence, self-discovery, personal growth, navigating young adulthood

Category 4: Music, Art & Creative Expression

This category celebrates artistic creativity, focusing on music, performance, and other forms of artistic expression. These stories highlight the struggles and triumphs of artists, musicians, and other creatives as they pursue their passions.

- Examples: Nailed It!, Money Heist: From Tokyo to Berlin, Rhyme & Reason
- Common Themes: Music, visual arts, creative struggles, passion for the arts

Category 5: Stories of Connection & Discovery

This category features narratives that emphasize human relationships, travel, and exploration. The characters in these stories embark on journeys—both physical and emotional—that broaden their perspectives and lead to meaningful discoveries.

- Examples: My Little Pony: A New Generation, Sankofa, The Great British Baking Show
- Common Themes: Exploration, adventure, self-realization, relationships, transformative experiences

Category 6: Life Stories & Documentary Narratives

This category includes real-life stories, biographical narratives, and documentary-style content. These titles often offer insight into societal issues, personal struggles, or historical events, providing a deeper understanding of real-world experiences.

- Examples: Dick Johnson Is Dead, Intrusion, Jailbirds New Orleans
- Common Themes: Real-life experiences, societal themes, deep reflection, thought-provoking narratives

7. Recommendation System

Our recommendation system personalizes movie and TV show suggestions using TF-IDF vectorization, word embeddings and cosine similarity. Given user inputs—genre ("Comedy"), type ("Movie"), actor ("Tom"), sentiment ("funny"), and keyword ("vacation")—we generate a

query vector, transform it using TF-IDF, and compare it with the Netflix dataset. The system selects the top five matches: Jim Gaffigan: Cinco, Tom Segura: Completely Normal, Jenny Slate: Stage Fright, Trigger Warning with Killer Mike, and Tom Segura: Ball Hog, all under "Stand-Up Comedy." Jim Gaffigan: Cincoscores highest at 0.2534, making it the best match, with the rest ranked by similarity

```
Enter your preferred genre (e.g., 'Comedy'): comedy
Enter your preferred type (e.g., 'Movie' or 'TV Show'): movie
Enter an actor/actress name (optional): tom
Enter sentiment words (e.g., 'funny', 'heartwarming'): funny
Enter additional keywords (optional): vacation

Recommended Titles (TF-IDF):
                              title  \
5639              Jim Gaffigan: Cinco
5379     Tom Segura: Completely Normal
3388         Jenny Slate: Stage Fright
4184  Trigger Warning with Killer Mike
2782              Tom Segura: Ball Hog

                                  listed_in     type
5639                          Stand-Up Comedy    Movie
5379                          Stand-Up Comedy    Movie
3388           Documentaries, Stand-Up Comedy    Movie
4184  Docuseries, Stand-Up Comedy & Talk Shows  TV Show
2782                          Stand-Up Comedy    Movie

Similarity Scores:
[0.25348861 0.19996897 0.1802892  0.17971545 0.17850016]
```

8. Conclusion & Chatbot

We developed a Recommendation System using the Netflix dataset and integrated it into a Chatbot for intuitive, personalized movie and TV show suggestions. Starting with data cleaning and preprocessing, we tokenized the description field and applied TF-IDF vectorization to convert text into numerical vectors. Using Cosine Similarity, we matched user queries with the dataset to filter the most relevant recommendations.

The Chatbot transforms this system into an interactive experience, allowing users to input genre, type, actor, sentiment keywords, and additional descriptions. It then processes the input, retrieves top-matching content, and presents structured recommendations with title, category, type, and similarity scores.

In the left example, the user searches “family”, and the system recommends stand-up comedy specials like Tracy Morgan: Staying Alive. The right example shows the Chatbot guiding the

user by asking for favorite actors (Ama Qamata) and genres (Comedy) before delivering tailored suggestions. This system streamlines content discovery through text processing, vectorization, similarity scoring, and conversational AI.