YouTube Suggestion Engine

Introduction

This project aims to explore and develop various recommender systems using different methodologies. Utilizing the YouTube dataset from Kaggle, our system recommends the top 10 most similar videos based on the title of a video watched or passed. We experimented with several approaches to build these recommender models. Initially, we employed an unsupervised learning method using Latent Dirichlet Allocation (LDA) for recommendations. Next, we built a content-based recommender system using TF-IDF embeddings and cosine similarity, which was further refined through additional data preprocessing and feature selection. We also experimented with BERT (Bidirectional Encoder Representations from Transformers) to enhance word embeddings. Finally, to improve recommendation relevance, we incorporated user feedback using a feedback-based recommendation system inspired by the Rocchio algorithm, which allows for precision adjustment based on user input.

Dataset

The YouTube dataset comprises approximately 40,881 videos from YouTube, with 16 columns containing information such as video_id, title, channel_title, category, description, and tags.

Recommendations Using LDA (Latent Dirichlet Allocation)

LDA is an unsupervised NLP model for topic discovery within documents, functioning similarly to clustering in numerical data. It reveals hidden patterns in text collections. In this project, LDA is applied to recommend videos based on their similarity to a given input video. The development process involves:

1. **Creating a New Dataframe:** The original dataframe is cleaned by retaining only the 'title', 'channel_title', 'tags', and 'description' columns. These columns are combined into a single 'overview' column, resulting in a dataframe with "title" and "overview" columns.
2. **Text Preprocessing:** Textual data is preprocessed in four stages: (i) removing non-English words, (ii) eliminating stop words, (iii) building bigrams, and (iv) lemmatization.
3. **Initializing the LDA Model:** LDA operates under four assumptions: (i) documents are bags of words, (ii) stop words carry no topic information, (iii) the number of topics (k) is known beforehand, and (iv) all topic assignments except for the word in question are accurate. The model is set with k = 10 to generate top 10 recommendations, identifying and ranking keywords in ten topics based on probability scores.
4. **Building the Recommender System:** The recommender system uses probability scores to classify input titles into topics, treating it as a classification problem. Videos with the highest scores are recommended to the user.

Content-Based Recommender System Using TF-IDF Embeddings

This system recommends videos based on similarities in descriptions and tags. The process includes:

1. **Data Preparation:** Similar to the LDA model, the data is cleaned and an 'overview' column is created.
2. **TF-IDF Vectorization:** The 'overview' column is transformed into a matrix of TF-IDF vectors using TfidfVectorizer.
3. **Cosine Similarity:** Video titles are indexed, and the vectors are compared using cosine similarity to find the most similar videos.
4. **Recommendation:** The model uses cosine similarity to rank videos and recommend the top 10 most similar ones based on user input.

Video Statistics-Based Adjustments

To simulate real-world recommendations, we incorporate video statistics by defining a similarity score threshold (0.05). Videos with scores above this threshold are ranked by view count, ensuring more relevant recommendations.

Content-Based Recommender System Using BERT Embeddings

BERT uses bidirectional training of transformers to model language, resulting in better recommendations compared to TF-IDF embeddings. BERT captures word context more effectively, enhancing recommendation quality.

Content-Based Recommendations Including User Feedback

Our final model incorporates user feedback into the TF-IDF-based recommender system, inspired by the Rocchio algorithm. This model adjusts recommendations based on user feedback, allowing for precision tuning. User feedback modifies the 'tags' through the following formula:

Augmented Tags = (0.8) * R - (0.1) * NR

where R represents the sum of embeddings for relevant documents and NR for non-relevant ones. This feedback loop enhances the precision and relevance of recommendations.

Evaluation

To evaluate recommendation quality, we used manual inspection and category-based metrics. We randomly selected 100 video titles and assessed recommendations from each algorithm based on category similarity. The relevance scores were 78% for TF-IDF, over 90% for LDA, and around 88% for BERT, indicating satisfactory relevance across all systems.

Conclusions

This project explored different recommender systems using the YouTube dataset from Kaggle. Outputs varied across models, with BERT providing recommendations based on genre similarity and TF-IDF focusing on same-artist recommendations. BERT's contextual understanding led to more relevant recommendations compared to TF-IDF.

Discussions

Limitations include the use of a limited dataset (popular YouTube videos from California) and the static nature of the dataset. Future improvements could involve training on a larger, dynamic dataset and using an ensemble approach to combine multiple recommender systems for more accurate recommendations.

Output of the systems for the input : **'Eminem - Walk On Water (Audio) ft. Beyoncé'**

LDA based Recommendation system :

```
recommend_by_title('Eminem - Walk On Water (Audio) ft. Beyoncé', df)
```

```
['Critical Role | Campaign 2 Episode 9',
 'Steam Code How To Get Free And Easily',
 "Marvel Studios' Black Panther - Warriors Of Wakanda",
 "Ufc 223: Khabib Nurmagomedov Reacts To Tony Ferguson'S Injury, Max Holloway Stepping In",
 'If He Were In Max Holloway'S Spot, Khabib Nurmagomedov Says He Wouldn'T Have Accepted Ufc 223 Fight
 "Snooki Explains Why She Fears Her Marriage Is Over On 'Jersey Shore: Family Vacation' (Exclusive)",
 'Pharmarusical (Season 10)',
 "Dc'S Legends Of Tomorrow 3X16 Promo I, Ava (Hd) Season 3 Episode 16 Promo",
 'John Mayer On Andy Cohen'S Annoying Habit | Wwhl',
 'Prank Hilarant – Youtube Hero #6']
```

Content based Recommendation system using tf-idf embeddings :

```
# Testing the recommendation system
give_rec('Eminem - Walk On Water (Audio) ft. Beyoncé')
```

```
1226     Walk On Water/Stan/Love The Way You Lie (Medle...
8020     Eminem - Walk On Water (Official Video) ft. Be...
5068                          Eminem - Untouchable (Audio)
6396                 Eminem - River (Audio) ft. Ed Sheeran
27728                                      Eminem - Framed
23871    Eminem - River (Behind the Scenes) ft. Ed Sheeran
7236                         Eminem - River ft. Ed Sheeran
6894                      Eminem - Believe (Official Audio)
265      Eminem Performs 'Walk On Water' | MTV EMAs 201...
7734     Eminem – River (Lyrics / Lyric Video) ft. Ed S...
Name: title, dtype: object
```

The output after consider the video statistics based adjustments (views):

```
startsearch()

Enter your search:
Eminem - Walk On Water (Audio) ft. Beyoncé
1 Eminem - Walk On Water (Official Video) ft. Beyoncé by EminemVEVO
2 Eminem - Walk On Water (Audio) ft. Beyoncé by EminemVEVO
3 Eminem Performs 'Walk On Water' | MTV EMAs 2017 | Live Performance by MTV International
4 Walk On Water/Stan/Love The Way You Lie (Medley/Live From Saturday Night Live/2017) by EminemVEVO
5 30 Seconds to Mars - Walk on Water (R3hab Remix) by Proximity
6 EMINEM Coachella 2018 (Full Live Performance) [Dr. Dre, 50 Cent & 2Pac] by Hip-Hop Universe
7 Eminem - River (Audio) ft. Ed Sheeran by EminemVEVO
8 Eminem - Untouchable (Audio) by EminemVEVO
9 Eminem - Nowhere Fast (Extended/Audio) ft. Kehlani by EminemVEVO
10 Eminem - Believe (Official Audio) by The Best for Loyalty
```

Content based Recommendation System using BERT embeddings:

```
[ ]  # Testing the recommendation system
     give_rec('Eminem - Walk On Water (Audio) ft. Beyoncé')
```

```
27728                                      Eminem - Framed
23871    Eminem - River (Behind the Scenes) ft. Ed Sheeran
21667      DJ Khaled ft. JAY Z, Future & Beyoncé - Top Off
21349    DJ Khaled - Top Off (Ft. JAY Z, Future & Beyonce)
6391     G-Eazy - No Limit REMIX (Audio) ft. A$AP Rocky...
33765    Enrique Iglesias - MOVE TO MIAMI (Official Vid...
647                      Remy Ma - Wake Me Up ft. Lil' Kim
7196     G-Eazy - No Limit REMIX ft. A$AP Rocky, Cardi ...
32323              Joyner Lucas & Chris Brown - I Don't Die
32707    Enrique Iglesias, Pitbull - Move To Miami (Lyr...
Name: title, dtype: object
```

: