YouTube Comment Analyzer:

Analyzing Comments for Genre Classification

BAX 422 Data Design and Representation Final Project

Group 7

Kumar Kishalaya

Yuna Kim

Omkar V. Shirigannavar

Table Of Contents

Executive Summary

Our project seeks to enhance the effectiveness of YouTube's video recommendation system by integrating data derived from YouTube comments. A key limitation is that current systems might not fully understand the content's context or genre, especially for newly uploaded videos with limited interaction data or for videos whose metadata is not descriptive enough.

This is why we extracted a comprehensive amount of primary comments from YouTube videos belonging to different genres, using Selenium and BeautifulSoup, followed by storing this data in an SQL database. It consisted of 500 rows for different videos, with each row containing all the primary comments associated with that video, as a list. We then use this dataset in our ML Project to then find insights from the data that has been collected and to see if comments are viable to group different types of videos together.

This project would use user-generated content to gain a deeper understanding of the audience's preference and market dynamics. This would enable YouTube to make data-driven improvements that would improve user engagement and drive growth.

Background and Context

Current Youtube users often struggle to discover new content that aligns with their interests. Traditional content recommendation algorithms rely on titles, descriptions, and tags along with user information data (views, likes, and watch history). However, these sources may not capture the full context or nuanced user sentiment towards the content.

By scraping the comments section of YouTube videos, a wealth of new user-generated content can be leveraged. These comments often contain viewers' opinions, discussions and reactions that provide insights on the contents of the video.

Data Sources and Description of Web Scraping Methods

Our project was based around the YouTube platform to extract comments from videos of specific genres. We ignored YouTube short content and only focused on the long form traditional videos of four genres - Sports, Documentaries, Movies and Lo-Fi. These genres had further keyword subcategories within them for more specificity.

Genre	Keywords
Sports	Football compilations, hockey compilations, cricket compilations, athletics top moments, tennis top moments
Documentary	Nature documentaries, History documentaries, Science documentaries, True crime documentaries, Technology documentaries
Movie	The Ultimate 2024 Oscar Trailers, Action movie trailers, 21st century best movies, Marvel movie trailers, Netflix original series
Lo-fi	Lo-fi hip-hop beats, Lo-fi chillhop mixes, Lo-fi study music, Lo-fi jazz playlists, Lo-fi ambient music

Search and Scraping Workflow

We first started by navigating YouTube for videos corresponding to specific combinations of genres and keywords followed by collecting the comments from each video. Here is a rundown of the steps involved:

1) Keyword Search on YouTube:

   First we leverage Selenium to perform searches on predefined keywords. Our goal here was to simulate the actions of a user interacting with and searching for content on YouTube, for which Selenium is ideal. After the search we were able to narrow down and retrieve videos that were most relevant to our specified keywords.

2) Comment Extraction with Scroll Limit:

   Given the extensive nature of comment sections on YouTube videos, we implemented a cap on how many comments we gather from each video. By doing this, we focus on a manageable subset of comment data rather than having discrepancies with some videos containing over 1,000 comments and others not having as much. This reduced the complexity and computation power that would have been required had we collected all comment data.

3) HTML File Generation per Video:

   Upon locating the video of interest, we utilized Selenium to scrape the comments from the comment section. In order to organize and enhance usability of the scraped data later in SQL, we generated the HTML files for each video.The filename is a combination of genre, keyword and video ID.

4) Duplicate Prevention:

   To ensure the integrity of our data and avoid duplication, we implemented measures within our code to avoid processing a video more than once. Before scraping we first check to determine if that video has already been analyzed. This precautionary step prevented redundant scraping efforts and maintained the accuracy of our dataset.

SQL Database Integration

To streamline the management and analysis of the scraped YouTube comments data, we created and stored the scraped data into SQL database. Here is a breakdown of how we went about structuring our database:

1) Mapping Genre to Category ID:

   To bring uniformity to our database, we linked each video's genre to a specific category_id. The mapping involved assigning unique identifiers to genres, to make it easier to perform genre-specific queries.

2) HTML File Processing and Data Insertion:

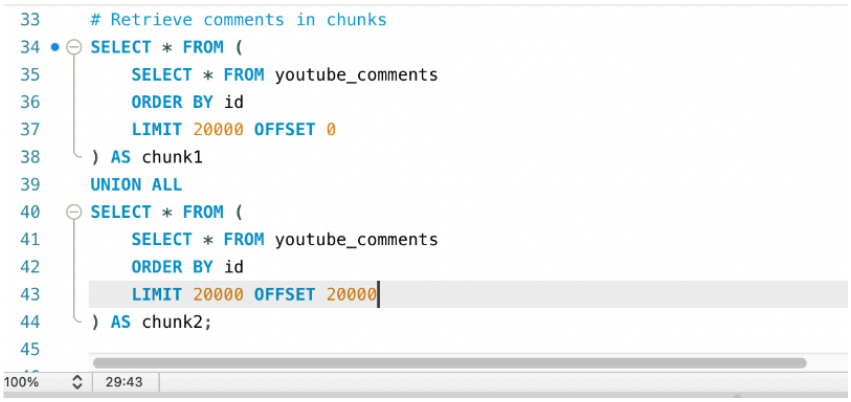
   Upon parsing each HTML file containing scraped comments, we extracted comment text, likes, genre, keyword, and video ID. Using BeautifulSoup, we processed the files to extract critical information such as comment text, number of likes per comment, genre, etc. which were then stored in an SQL table.

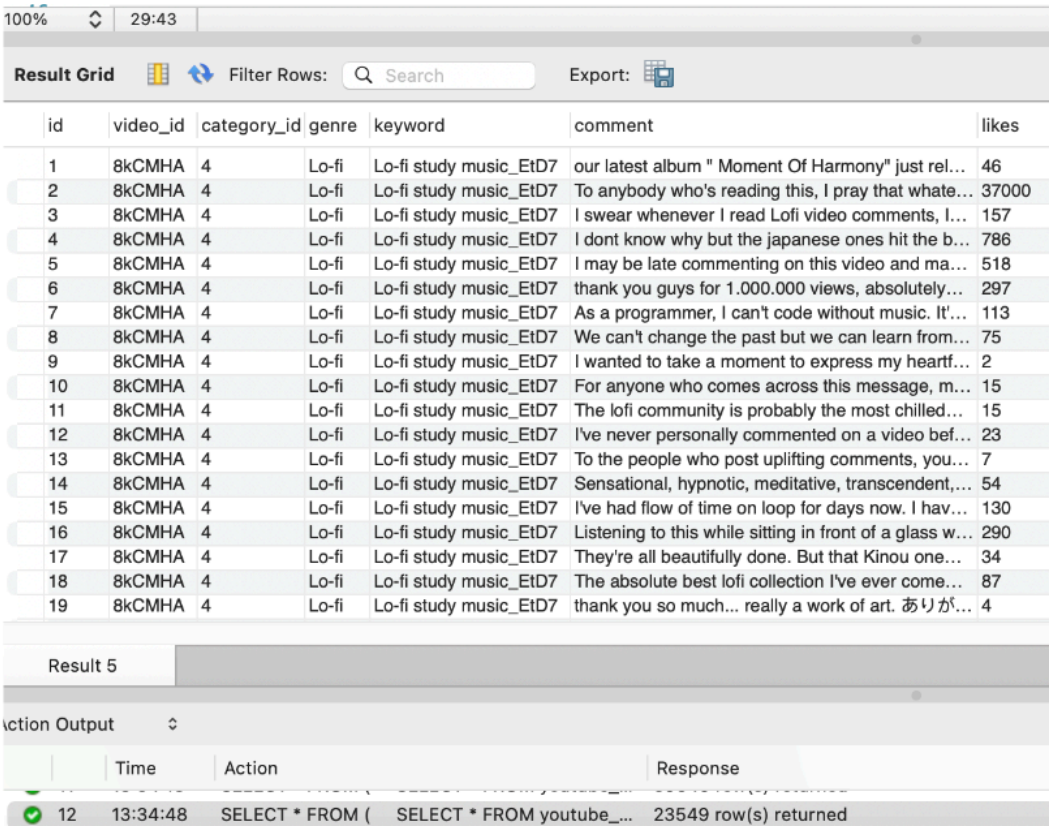
3) Exporting Data to CSV:

   Once all scraped comments were successfully stored in the SQL database, the final step was to export this data into a CSV file. This CSV would then be used for further analysis in our ML project.

SQL data dictionary

- ID: Unique identifier for each comment of a video
- Video_id : Identifier for videos to link to specific videos in a database
- Category_id : Category identifier used to classify videos into different genre.
- Genre : Genre of video content
- Keyword : Search terms
- Comment : User-generated comment for each videos
- Likes : Number of likes for each comment





100% 29:43

Result Grid Filter Rows: Search Export:

id	video_id	category_id	genre	keyword	comment	likes
1	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	our latest album " Moment Of Harmony" just rel…	46
2	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	To anybody who's reading this, I pray that whate…	37000
3	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	I swear whenever I read Lofi video comments, I…	157
4	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	I dont know why but the japanese ones hit the b…	786
5	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	I may be late commenting on this video and ma…	518
6	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	thank you guys for 1.000.000 views, absolutely…	297
7	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	As a programmer, I can't code without music. It'…	113
8	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	We can't change the past but we can learn from…	75
9	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	I wanted to take a moment to express my heartf…	2
10	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	For anyone who comes across this message, m…	15
11	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	The lofi community is probably the most chilled…	15
12	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	I've never personally commented on a video bef…	23
13	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	To the people who post uplifting comments, you…	7
14	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	Sensational, hypnotic, meditative, transcendent,…	54
15	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	I've had flow of time on loop for days now. I hav…	130
16	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	Listening to this while sitting in front of a glass w…	290
17	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	They're all beautifully done. But that Kinou one…	34
18	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	The absolute best lofi collection I've ever come…	87
19	8kCMHA	4	Lo-fi	Lo-fi study music_EtD7	thank you so much... really a work of art. ありが…	4

Result 5

ction Output

	Time	Action		Response
12	13:34:48	SELECT * FROM (	SELECT * FROM youtube_...	23549 row(s) returned

Recommendations and Business Value provided

Scraping YouTube comments to form a dataset for classifying videos into different genres addresses various business questions and objectives. One of the major themes for our project is improving the current recommendation system that YouTube has in place. Comments have real-world user generated reactions and interactions with the video that can enhance machine learning models. Including these comments as another factor can lead to more nuanced and personalized content, suggestions, improving user engagement and satisfaction.

It could also help uncover content that may have been overlooked by traditional metadata-based recommendation methods. Looking at user interactions and sentiments expressed in comments more in depth, we can identify niche content that resonates with specific audience segments. This understanding empowers YouTube to diversify its content recommendations, catering to a broader range of user preferences and interests.

Summary and Conclusions

This project has effectively gathered a substantial volume of comments spanning diverse YouTube video genres, including Sports, Documentaries, Movies, and Lo-Fi. By leveraging user-generated content from comments, this project has demonstrably opened doors to enhance YouTube's video recommendation system.

The rich data set offers a valuable resource for training machine learning models. These models can then be incorporated into the recommendation system, leading to more nuanced and personalized content suggestions for users. Future research can delve deeper by exploring sentiment analysis and topic modeling of these comments, providing even more insights into user preferences and content dynamics. Overall, this project paves the way for integrating user-generated content as a key factor in future recommendation algorithms, ultimately fostering a more personalized and engaging user experience on YouTube.