YouTube Data Collection And Analysis For Indian Data

23ed September 2024

Project Overview :

The YouTube Data Collection and Analysis project aims to gather data on YouTube videos using Google Cloud Console-generated YouTube APLs. This project will focus on collecting data about trending videos in India and conducting an in-depth analysis to identify factors contributing to their popularity, such as video category, view count, like count, comment count, and video length. The analysis will provide valuable insights into user behaviour and content preferences, helping content creators and marketers make data-driven decisions.

Objectives :

- **Data Collection:** Use the YouTube Data API to extract data on trending videos, including video metadata, statistics, and user engagement metrics.
- **Data Analysis**: Analyze the collected data to uncover patterns and engagement (views, likes, comments).
- **Predictive Modeling:** Develop predictive models to forecast video popularity based on historical data, helping creators optimise their content strategies.
- **Visualisation:** Create clear and insightful visualisations (graphs, charts) to represent the trends and patterns discovered during the analysis.

Dataset Description:

- **video_id:** Unique identifier for each video.

- **title:** Title of the video.

- **description:** Description provided by the video uploader.

- **published_at:** Date and time when the video was published.

- **channel_id:** Unique identifier for the YouTube channel.

- **channel_titel:** Name of the YouTube channel.

- **category_id:** Numerical ID representing the category of the video.

- **tags:** A list of tags associated with the video.

- **duration:** Quality of the video (e.g., HD).

- **caption:** Boolean value indicating if captions are available.

- **view_count:** Number of views the video has received.

- **like_count:** Number of likes the video has received.

- **dislike_count:** Number of dislikes the video has received.

- **favorite_count:** Number of times the video has been favourited.

- **comment_count:** Number of comments the video has received.

Scope :

The project will involve the following steps:

- **API Integration:** Set up the YpuTube Data API using Google Cloud Console and obtain access keys for data extraction.
- **Data Collection:** Collect data on trending videos, including:
- Video Title
- Channel Name
- Category
- Published Date
- View Count
- Like Count
- Comment Count
- Video Lenght
- Tags
- Description
- **Data Preprocessing:** Clean and preprocess the data by handling missing or duplicate entries and formatting the dataset for analysis.
- **Exploratory Data Analysis (EDA):** Perform EDA to understand the key factors influencing video trends and user engagement. Analyze the following:
    - Most popular categories
    - Relationship between video length and view count
    - Impact of publication time on popularity

- Engagement metrics like likes and comments
- **Machine Learning Models:** Use regression or classification models to predict video popularity based on factors like category, tags, and user interaction data.
- **Visualization:** Create dashboards or interactive visualisations to present findings in a user-friendly manner.

Deliverables:

- **Data Collection:** A cleaned dataset of trending YouTube videos in India.
- **Exploratory Data Analysis:** Insights and visualizations on trends and factors affecting video popularity.
- **Predictive Models:** Machine Learning models predicting video success based on collected data.
- **Final Report:** A report summarizing insights, analysis, and recommendations for content creators.

Tools And Technologies:

- Programming Language: Python
- API: YouTube Data APL v3 (Google Cloud Console)
- Data Processing: Pandas, NumPy
- Visualisation: Matplotlib, Seaborn, Plotly, Dash (for interactive dashboards)
- Machine Learning: Scikit-learn

Evaluation Criteria:

- Accuracy of Data Collection: Successful integration with YouTube API and correct data extraction.
- Data Analysis Quality: Depth of insights derived from EDA and relevance to understanding YouTube trends.
- Model Performance: Effectiveness of predictive models in forecasting video popularity.
- Visualisation Clarity: Quality and clarity of visual representations and dashboards.