YouTube Trending Video Analytics

Introduction

In today's digital age, YouTube has become one of the most influential platforms for video content and audience engagement. This project aims to uncover patterns in trending YouTube videos by analyzing datasets from different regions (India, US, Canada, UK). The goal is to identify the most engaging genres, sentiment trends, regional differences, and category performance using data science and visualization tools.

Abstract

This project analyzes YouTube trending data to uncover content trends, viewer preferences, and sentiment patterns. It covers data cleaning, sentiment analysis, SQL queries, time-series insights, and a Power BI dashboard to understand what content trends, for how long, and across which regions.

Tools Used

- **Python:** Data cleaning, preprocessing, sentiment analysis (TextBlob), and visualizations (Matplotlib, Seaborn)
- **SQL (Oracle 11g Express Edition):** Aggregation, ranking, and category-wise analysis
- **Power BI:** Interactive dashboards for region-wise comparisons and top category analysis

---

Steps Involved in Building the Project

1. **Dataset Collection**
   Downloaded trending YouTube video datasets from Kaggle (IN, US, CA, GB) in CSV format along with category JSON files for mapping category names.

2. **Data Cleaning and Merging**
   - Removed null and duplicate rows
   - Standardized column names
   - Merged video data with category metadata for all countries
   - Filtered relevant columns (title, tags, views, likes, dislikes, etc.)
   - Saved as youtube_trending_filtered.csv (20K+ rows)

3. **Sentiment Analysis**
    - Performed sentiment polarity scoring on video titles and tags using TextBlob
    - Classified sentiments as Positive, Neutral, or Negative
    - Added sentiment columns to the dataset for visualization

4. **SQL Analysis**
   To derive regional and sentiment-based insights, SQL queries were performed on the cleaned dataset stored in Oracle 11g:

   Exploratory Queries:
    - Fetched top 10 rows to verify structure.
    - Counted total videos per region (IN, US, CA, GB).
    - Identified top 10 most viewed videos globally.

   Analytical Queries:
    - Calculated sentiment distribution (Positive, Neutral, Negative).
    - Computed average likes per region to assess viewer engagement. Ranked top 5 categories by average views using GROUP BY.

5. **Time-Series and Trend Duration Analysis:**
    - Used Python to calculate and visualize how long videos stayed on the trending list
    - Created line plots showing daily trends, country-wise peaks, and trending duration by category

6. **Dashboard Creation in Power BI**
   Built a 3-page interactive dashboard:
    - **Executive Summary:** Shows KPIs like total views, top regions, and video count.
    - **Genre & Sentiment:** Visualizes top categories and sentiment distribution across regions.
    - **Time Series:** Highlights trending patterns and duration over time.

Added slicers for region, category, and sentiment for easy filtering.

Conclusion

The project successfully uncovered insights about content popularity across different countries and how sentiments and video categories influence trending patterns. The combination of Python, SQL, and Power BI allowed for a complete data pipeline — from preprocessing to storytelling. This project demonstrates the power of data analytics in understanding digital content trends and prepares me for real-world analytical challenges.