Project G2:
YOUTUBE TRENDS

Analysing YouTube trends and training a video category prediction model

TEAM:

Timofei Šinšakov (Gr 7)

Natalja Frantikova (Gr 4)

Andrei Potrebin (Gr 2)

Project repository in GitHub

Identifying business goals

Background

YouTube currently stands as one of the most popular platforms for video content consumption. YouTube's "trending" section highlights videos that have captured widespread attention in a specific region. Considering the immense popularity of the platform, and the fact that users' activity is what defines what videos will be trending, conclusions on broader societal interests could be made by looking at the YouTube trends. Analyzing the data from two major markets, the US and the UK, meaningful insights into media consumption and culture could be gathered. Such a project would have many benefits, ranging from what topics are currently captivating public attention of a country's media users, to helping creators align their content strategies with relevant topics. This project will seek to derive insights into viewing habits, content characteristics, and trends over time.

A note on the target audience this project benefits

The project does not benefit a business (company). Instead, it primarily benefits content creators, marketers, cultural analysts, and researchers of media trends.

Business goals

1) Identify and compare trends in video categories and topics between the US and the UK to highlight regional differences in audience preferences.

2) Illustrate the evolution of popular themes on YouTube across the period from august 2020 to april 2024.

3) Introduce a simple predictive model that would tell a user what category label a video aligns with best based on the video's metadata.

Business success criteria

Goal 1 is achieved if the analysis identifies at least three important differences and three similarities in video categories and topics between the US and the UK, presenting the findings with visualizations and descriptions.

Goal 2 is achieved if the evolution of popular topics among YouTube trends is illustrated with at least three distinct and informative visualizations over a 44-month period.

Goal 3 is achieved if the predictive model predicts a video's category correctly in at least 80% of cases during testing.

Assessing your situation

Inventory of resources

Data resources include two (identically structured) datasets of (1) trending YouTube videos from the UK and (2) trending YouTube videos from the US, each having more than 45 000 unique videos. Tool inventory consists of Jupyter Notebook, python and its libraries for data mining, machine learning and visualization tasks.

Requirements, assumptions, and constraints

Datasets need to be prepared for analysis, ensuring that there are no rows or cell values that could potentially inhibit the project (like duplicates or NaN values).

It is assumed that captured video attributes are accurate in the dataset, and that the video trends are representative of larger audience behaviors from their respective countries.

Data constraints: trending videos only cover a 44-month time period up until the april of 2024. This means that no reliable claims about long-term trends can be made. It also has some effect on the relevance of project findings since latest trends are not in the dataset.

Risks and contingencies

The prediction model may fail to achieve desired accuracy. Contingency: try to enhance performance by using techniques like hyperparameter tuning, feature engineering, or additional data acquisition.

Some values in the dataset might lead to incorrect generalizations and observations during the analysis stage. Contingency: identifying values and keywords that are over-general on YouTube and adjusting the python code to ignore them.

Terminology

Trending video: a video that appears in YouTube's "Trending" section for a specific region.

Video category: the name of the general content type that the video is labeled with. It is chosen from a list of categories by the video's author when publishing the video.

Engagement: measures of audience interaction, including views, likes, dislikes, and comments.

Costs and benefits

A significant amount of time will be required for data cleaning. It will benefit the accuracy of the analysis and the reliability of results.

Much time and computational power will be needed for trying out several models and tuning hyperparameters on them. The benefit is the guarantee that many potential prediction models will be trained and tested.

Defining your data-mining goals

Data-mining goals

1. Train and assess a machine learning model that predicts video content category based on title, tags, and description.

2. Conduct a comparative analysis of content of trending videos from US and UK based on category, title, description, tags, publishing channel, and audience engagement.

3. Identify tendencies of how trends and general audience's interests were changing over time.

Data-mining success criteria

Goal 1 is accomplished if a prediction model is trained and its overall accuracy is at least 80%.

Goal 2 is accomplished if statistical tests show a p-value < 0.05 for the prevalence of certain content categories among trending videos in one country compared to the other.

Goal 3 is accomplished if temporal changes in trending topics are identified using time series analysis with a confidence interval of 95%.

1. Gathering Data

Data Requirements and Availability

The datasets for this project were sourced from Kaggle and include two datasets covering trending YouTube videos from the USA and the UK between August 4, 2020, and April 14, 2024. The datasets contain the following fields:

1. **video_id**
   - **Description:** A unique identifier assigned to each YouTube video.
   - **Relevance:** Not directly useful for this project, as it does not contribute to the prediction model or comparative analysis.
2. **title**
   - **Description:** The title of the video as displayed on YouTube.
   - **Relevance:** A key input for the machine learning model. Titles often contain keywords that hint at the video's category and content focus.
3. **publishedAt**
   - **Description:** The date and time when the video was published on YouTube.
   - **Relevance:** May provide insights into publishing trends (e.g., time of year or day of the week).
4. **channelId**
   - **Description:** A unique identifier for the channel that uploaded the video.
   - **Relevance:** Not useful for the project's objectives and excluded from analysis.
5. **channelTitle**
   - **Description:** The name of the YouTube channel that published the video.
   - **Relevance:** Potentially useful for understanding brand consistency or identifying popular creators, though not included in model training.

6. **categoryId**

    - **Description:** A numerical ID representing the video's category. A JSON mapping file translates these IDs into readable category names.
    - **Relevance:** The target variable for the prediction model. Ensuring proper mapping and inclusion in the dataset is critical.

7. **trending_date**

    - **Description:** The date when the video was trending on YouTube.
    - **Relevance:** Useful for tracking the popularity lifecycle of videos. May help correlate trends with real-world events.

8. **tags**

    - **Description:** A list of user-provided tags associated with the video.
    - **Relevance:** Essential for prediction modeling as tags often summarize the video's content. Requires preprocessing for consistency.

9. **view_count**

    - **Description:** Total views the video received up to the time it trended.
    - **Relevance:** Provides a measure of popularity but not directly used in predictions.

10. **likes**

    - **Description:** Total number of likes received up to the time the video trended.
    - **Relevance:** Indicates audience engagement, useful for comparative analysis between the USA and the UK.

11. **dislikes**

    - **Description:** Total number of dislikes received. This field is set to zero for most records due to YouTube's disabling of the dislike counter.
    - **Relevance:** They are considered irrelevant and are excluded from the analysis since YouTube disabled their display.

12. **comment_count**

    - **Description:** Total number of comments on the video at the time it trended.
    - **Relevance:** May indicate audience interaction levels but not directly relevant to prediction tasks.

13. **thumbnail_link**

    - **Description:** URL of the video thumbnail.
    - **Relevance:** Not relevant for this project as thumbnails are visual and not textual or categorical data.

14. **comments_disabled**

    - **Description:** A boolean field indicating whether comments are disabled for the video.
    - **Relevance:** Useful for understanding engagement strategies but not directly part of the prediction model. Presumably not needed in the analysis.

15. **ratings_disabled**

    - **Description:** A boolean field indicating whether ratings (likes and dislikes) are disabled for the video.
    - **Relevance:** Helpful for understanding restrictions on audience interaction but not included in model predictions. Presumably not needed in the analysis.

16. **description**

    - **Description:** The text description provided by the uploader for the video.
    - **Relevance:** A critical input for the machine learning model. Often contains detailed context about the video's content.

17. **Derived Field: days**

    - **Description:** A calculated field indicating the total number of days a video remained trending.
    - **Relevance:** This derived metric will help determine a video's true popularity and contribute to identifying top trends.

Selection Criteria

For the purpose of this project, the following decisions were made:

1. **Irrelevant fields:** video_id, channelId, dislikes, comments_disabled, ratings_disabled were deemed unnecessary. dislikes is particularly redundant as YouTube has disabled the counter.
2. **Focus Fields:** The prediction model will rely on title, tags, and description to predict the video category.
3. **Additional Field:** A new column, days, will be derived to quantify how long a video remained trending.

2. Describing Data

Data Overview

- **Source:** Datasets were downloaded from Kaggle.
- **Format:** CSV files, which are compatible with data analysis tools.
- **Fields and Records:** Each dataset contains 16 fields, and the total record count is 268787 for the USA and 268791 for the UK.

Suitability

The datasets include all fields required for the machine learning objectives. The time frame (2020–2024) is comprehensive enough to cover variations in trends and external events.

3. Exploring Data

Variable Analysis

A brief analysis has already been done in the description of the fields, but in short:

- **Text Fields** (title, tags, description): The most critical features for predicting the category of a video.
- **Temporal Features** (trending_date, days): Useful for analyzing seasonality and real-world event correlations.
- **Engagement Metrics** (view_count, likes, comment_count): Relevant for comparing trends but not for predictive modeling.

Hypotheses

- A minor difference in video themes between the US and UK are expected. Presumably, politics will be more common in the US, and sports in Britain..
- A strong correlation is expected between trending_date and real-world events, such as newsworthy occurrences.

Initial Observations

Duplicate entries are observed due to videos trending over multiple days. The days column will aggregate these occurrences to better evaluate true popularity.

4. Verifying Data Quality

- **Completeness:** All required fields are present. However, missing values are noted in tags, description, views_count, likes, dislikes, comment_count due to the disabled ratings and comments.
- **Consistency:** Duplicate entries (by video_id) are intentional, reflecting trends across multiple days.
- **Accuracy:** The data is accurate because it was taken from YouTube API.
- **Relevance:** The focus fields align with the project's goals.

Project plan

Goal 1: Train a Machine Learning Model to Predict Video Categories

Tasks:

1. **Data Preprocessing**:
    a. Combine, clean and vectorize data for text representation.
        i. Andrei - 1 h
        ii. Timofei - 2h
2. **Model Training**:
    a. Train a predictive model on the processed features.
        i. Andrei - 4 h
        ii. Timofei - 6 h
3. **Model Evaluation**:
    a. Test the model on unseen data.
        i. Timofei - 1h
    b. Calculate accuracy and classification metrics.
        i. Timofei - 5h
    c. Evaluate how well the model predicts video categories.
        i. Andrei - 3 h
        ii. Timofei - 3h
        iii. Natalja - 2 h

Goal 2: Compare Popular Videos from the US and GB

Tasks:

1. **Data Segmentation**:
    - Clean, deduplicate, and identify common videos and channels between datasets.
        i. Andrei - 1 h
2. **Feature Comparison**:
    - Analyze categories, channels, tags, and trending durations between US and UK.

        i. Andrei - 5 h
        ii. Natalja - 5 h
        iii. Timofei - 5h

3. **Visualization**:
    - Generate pie charts, histograms, word clouds, and daily comparisons to explore and present differences.
        i. Andrei - 4 h
4. **Statistical Analysis**:
    - Calculate percentages and frequencies for cross-country comparisons of videos, channels, and tags.
        i. Andrei - 6 h
        ii. Natalja - 5 h
        iii. Timofei - 4h

Goal 3: Investigate Popular Themes Over Time

Tasks:

1. **Temporal Analysis**:
    - Extract and analyze year and month-level data to track changes over time for views and categories.
        i. Natalja - 3 h
        ii. Andrei - 2 h
2. **Theme Research**:
    - Research the spikes in popularity of certain themes in plots and correlate the findings with real events.
        i. Natalja - 10 h
        ii. Andrei - 5 h
        iii. Timofei - 3h
3. **Visualization**:
    - Plot trends over time using line charts and heatmaps.
        i. Natalja - 2.5 h
    - Generate word clouds to visualize popular tags.
        i. Natalja - 2.5 h
        ii. Timofei - 1h

Tools and Methods:

- Data Processing: pandas
- Text Analysis: WordCloud
- Machine Learning: Scikit-learn
- Visualization: Matplotlib, Seaborn