CSP 571 - Data Preparation and Analysis
Project - Proposal & Outline
Spring 2024

Group members -
Bhavana Dontamsetti A20511041
Tanaya Vivek Mahakalkar A20518511

1. Project Proposal

1.1 Description of the project and research goal

YouTube has become a dominant platform for sharing and consuming video content, with millions of videos uploaded and viewed every day. The goal of this research is to understand the dynamics of YouTube trending videos and identify key factors that contribute to their success.

This project aims to conduct an exploratory data analysis (EDA) on YouTube trending videos to gain insights into the factors influencing video popularity. By examining trends in video categories, publishing times, and engagement metrics, the project seeks to uncover patterns and correlations that can help content creators and marketers optimize their strategies for reaching a wider audience.

1.2 Set of questions the project seeks to address

I. Which video categories are most likely to trend on YouTube?
II. What is the optimal duration for a video to trend categorized in a country?
III. Does the time of day or day of the week of publishing affect a video's likelihood of trending?
IV. Are there specific keywords or tags that are more likely to result in a video trending?
V. Can sentiment analysis be used to predict the likelihood of a video trending based on the sentiments expressed in its comments or title?
VI. Can clustering be used to identify niche content categories within YouTube that have a higher likelihood of producing trending videos?
VII. Are the same trends being followed even today?

1.3 Proposed methodology/approach

First, we will perform preprocessing of data, i.e cleaning the data and preparing it for analysis and predict some outcomes from the data and plot them on graph to derive predictive results for future

1.3.1 Data Collection

- Import data sources stated using R.
- Firstly, analyze a 5 year old dataset which has 10 csv files representing 10 countries with 16 columns each.
- Use Youtube API/ scrape data for analyzing if the trends today are still the same.

1.3.2 Data Preprocessing

- Clean the dataset by removing irrelevant columns and handling missing values.
- Perform text preprocessing on titles, comments and video description for sentiment analysis, like removing stopwords, and stemming or lemmatization.

1.3.3 Exploratory Data Analysis (EDA)

- Conduct statistical tests to determine if there are significant differences in sentiment scores and engagement metrics between clustering groups or trending and non-trending videos.
  We aim to analyze in the following manner:

  1. Country and category analysis
     - Explore the Variables relating to country
     - Analysis on languages used
     - Analysis on country and category relationship

  2. Sentiment Analysis:
     - Cleaning of Variables
     - Sentiment Distribution
     - Relationship Analysis of the sentiment and popularity

  3. Clustering Analysis:
     - Feature Engineering
     - Clustering Techniques
     - Cluster Visualization

1.4 Metrics for Measuring Analysis Results:

The following metrics will help quantify the results of the analysis at each step.

1. Exploratory Data Analysis: Utilize metrics like data distribution, central tendency measures, and data variability to understand the dataset's characteristics.

2. Sentiment Analysis: Measure the quality of sentiment distribution (positive, negative, neutral), average sentiment score.

3. Clustering Analysis:Evaluate clustering quality using metrics like silhouette score and inertia to assess the effectiveness of grouping videos based on common characteristics.

2. Project Outline

2.1 Literature review and related work

2.1.1 Data repository

1. https://www.kaggle.com/code/donyoe/exploring-youtube-trending-statistics-eda/in

2. Use scraped Dataset from youtube.(using YoutubeAPI)

2.1.2 Reference resources:

1. Youtube_Trending_Videos_Boosting_Machine_Learning_Results_Using_Exploratory_Data_Analysis - https://www.researchgate.net/publication/355576123_Youtube_Trending_Videos_Boosting_Machine_Learning_Results_Using_Exploratory_Data_Analysis

2. YouTube Data Analysis & Prediction of Views and Categories - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4076559

3. Trending Videos: Measurement and Analysis https://arxiv.org/pdf/1409.7733.pdf

2.1.3 Supplemental resources:

https://medium.com/@vijay_sundaram/analysing-popular-travel-youtubers-an-eda-project-using-youtube-video-data-in-python-298910922603

2.2 All data sources and reference data with descriptions

Dataset Type: Multivariate
Dataset Size: 10 documents categorized by country, each having 16 columns and around 20000 records each.
Countries chosen: Canada, USA, Great Britain, France, Denmark, Russia, Mexico, South Korea, India, Japan
Missing Values: Yes

Feature Descriptions:

- **video_id:** A unique identifier for each video on YouTube. (String)
- **title:** The title of the video. (String)
- **trending_date:** date when the video appeared on the trending list. (String)
- **channel_title:** name of the YouTube channel that uploaded the video. (String)
- **category_id:** category to which the video belongs. (Integer)
- **published_time:** This is the date and time when the video was originally published on YouTube. (DateTime)
- **tags:** Tags are keywords or phrases that describe the content of the video. (String)
- **views:** The number of views. (Integer)
- **likes:** The number of likes. (Integer)
- **dislikes:** The number of dislikes. (Integer)
- **comment_count:** The number of comments. (Integer)
- **thumbnail_link:** link to the thumbnail image that represents the video. (String)
- **comments_disable:** Whether comments have been disabled for the video. (Boolean)
- **ratings_disable:** Whether ratings have been disabled for the video. (Boolean)
- **video_error_or_removed:** whether there was an error with the video or if the video has been removed from YouTube. (Boolean)
- **description:** The description of the video. (String)

Category Types and respective ID:

category_data= [
(1, 'Film & Animation'),
(2, 'Autos & Vehicle'),
(10, 'Music'),
(15, 'Pets & Animals'),
(17, 'Sports'),
(19, 'Travel & Events'),
(20, 'Gaming'),
(22, 'People & Blogs'),
(23, 'Comedy'),
(24, 'Entertainment'),
(25, 'News & Politics'),
(26, 'Howto & Style'),
(27, 'Education'),
(28, 'Science & Technology'),
(30, 'Movies'),
(43, 'Shows'),
(29, 'Nonprofits & Activism')
]

2.3 Data processing and pipeline

1. Cleaning:
   - Handle missing values in numerical features like views, likes, comments using mean or median imputation.
   - Convert categorical variable category_id into dummy variables for modeling.
   - Normalize numerical features like views, likes, and comments to ensure that all features have the same scale.

2. Transformation:
   - Convert the 'trending_date' and 'published_time' columns to datetime format for easier manipulation.
   - Extract additional features from 'published_time', such as the day of the week or hour of the day the video was published.

3. Outlier Detection:
   - Use statistical methods like z-score, IQR to detect outliers in numerical features like views, likes, and comments. Consider removing outliers or transforming the data to reduce their impact on the analysis.

4. Sentiment Analysis and Clustering:
   - Perform sentiment analysis on the 'title', 'description', and 'tags' columns to extract sentiment scores

- Evaluate the clustering results to understand patterns and similarities among videos on similar characteristics such as views, likes, comments, and sentiment scores..

2.4 Data stylized facts

1. Distributional Analysis:
   - Conduct distributional analysis on numerical features like views, likes, comments, and sentiment scores to understand their distributions.
   - Visualize the distributions using histograms, box plots, or kernel density plots.

2. Clustering:
   - Use clustering algorithms to group videos based on similar characteristics.
   - Analyze the clusters to identify common traits among videos within each cluster.

2.5 Model selection

Feature Selection Requirements and Clustering:
- Select relevant features for the model based on their importance in predicting video trends like views, likes, comments, sentiment scores.
- Use techniques like feature importance from tree-based models or correlation analysis to select features.
- Implement clustering algorithms based to identify similarities.

2.6 Software packages

- Libraries/Packages: tibble, DT, knitr, tm, ggplot2, wordcloud, dplyr, fitdistrplus, plotly, plyr, textblob, cluster, Youtube API
- Softwares/Languages: RStudio, Jupyter Notebook, R , Python (for YoutubeAPI)