Team 28 Project Proposal

Team Member: Mitali Bante, Tracy Cui, David Ma, Ashutosh Rawat, Wei-Lun Tsai

Project Title: Trending Youtube Video Analysis

Background and Context:
As Youtube becomes more and more popular nowadays, our group is interested in exploring the top trending videos and learning more about various characteristics of popular videos. We found this dataset on Kaggle that is connected to Youtube API and contains a list of data on daily trending YouTube videos. This dataset includes data for several months and for different regions/countries. Also, it contains basic and relatable information about each video (for example, video title, tags, publish time, tags, views, likes and dislikes etc.). We would like to apply advanced data analysis and machine learning techniques to find more interesting stories about those trending videos. More specifically, some of the questions we are interested in figuring out are: why are those videos popular? What kinds of videos are popular? What factors affect how popular a video will be? How can we categorize those trending videos and make better recommendations for the audience?

Dataset:
This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, FR, RU, MX, KR, JP and IN regions (USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan and India respectively), with up to 200 listed trending videos per day. Each region's data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

The data contains the following information:

- **Video_ID:** Unique value for each video
- **Trending_date**: The date the video started trending (format: YY.DD.MM)
- **Title**: The title of the video
- **Channel_title:** The channel owner
- **Category_id:** Category of the video (musics, sports…) in the form of an integer
- **Publish_time:** Date and time the video was released (UTC)
- **Tags:** Keywords used to categorize the video
- **Views:** Number of views on the video
- **Likes:** Number of likes on the video
- **Dislikes:** Number of dislike on the video
- **Comment count:** Number of comments on the video
- **Thumbnail_link:** Preview image of the video
- **Comments_disabled:** Publisher choice if audience can comment or not (True or False)
- **Ratings_disabled:** Publisher choice if audience can rate (like or dislike) the video (True or False)
- **Video_error_or_removed**: Video existence (True or False)
- **Description:** Description/Advertisement of the video as uploaded by the channel

Proposed ML Techniques:
- Exploratory Data Analysis on the complete data to gain valuable insights and work on the pre-processing
- Analyse the trending youtube videos on country level, and see the trending categories over a time period
- Explore different supervised Machine Learning algorithms and implement multiple hyperparameter tuning techniques to predict the number of likes/ dislikes/ views and analyse factors responsible for the popularity
- Utilize NLP techniques to prediction the category type and build a recommendation system that will suggest similar videos based on the hashtags used