Capstone Project II proposal

Title:

YouTube video views in the US and UK: Exploratory data analysis, inferential statistics, and classification

Objective:

The objective of the project has three folds: exploratory analysis, inferential statistics, and classification. In the exploratory data analysis, behavioral patterns of youtube viewers, for example, number of youtube video views, likes, and dislikes in the US and UK are analyzed. In the inferential statistical analysis, distributions of number of views in the US and UK are examined based on the null hypothesis that there is no significant difference between the distribution of number of views in the US and UK. In classification, two machine learning models, logistic regression and random forest are applied to classify the US youtube videos and comparison of ML models are presented.

The project aims to answer the following questions:
(i) How are the top trending videos different in terms of views, categories, likes and dislikes?
(ii) Does there exist any seasonal pattern in some specific categories?
(iii) How are videos categories different in the US and UK?
(iv) Is the distribution of number of views in the US and that in the UK are significantly different?
(v) Can we classify the youtube videos as ‘extreme’, ‘high’, ‘medium’, and ‘low’ based on their number of views, category, day of the video is watched (say, weekends or weekdays), and month of the video watched?

The project will cover applications of data loading, data cleaning, exploratory data analysis, inferential statistics, and machine learning classification.

Data description:

The data source is https://www.kaggle.com/datasnaek/youtube. The dataset contains 6 files -

- 3 files on US viewers: UScomments.csv, USvideos.csv, US_category_id.json and
- 3 files on UK data: UKcomments.csv, UKvideos.csv, UK_category_id.json.
- USvidoes.csv has following 11 columns: ['video_id', 'title', 'channel_title', 'category_id', 'tags', 'views', 'likes', 'dislikes', 'comment_total', thumbnail_link', 'date'] where unique 'video_id's are 2364.
- UScomments.csv has 4 columns: ['video_id', 'comment_text', 'likes', 'replies'], where unique 'video_id's are 2266.
- US_category_id.json has [category_id, kind, etag, item_snippet], where unique category_id are 16.