You Tube Video Analytics

Spark

**YouTube Video Analytics**

**Problem Statement:**

YouTube uses few key factors like measuring user interactions; number of views, shares, comments, and likes etc. to rank the videos as the top trending on the platform. The top trending list of videos mostly belongs to movies, music, celebrities, reality shows, TV shows, etc. This dataset includes a few months of daily trending YouTube videos.

The dataset includes a category\_id field and its mapping is present in a JSON file.  
  
**Objective:** Use Spark features for data analysis and showing the valuable insights.

**Domain:** YouTube video analytics

**Analysis to be done:** Exploratory analysis to determine actionable insights

**Dataset:**

* USvideos.csv

video\_id,trending\_date,title,channel\_title,category\_id,publish\_time,tags,views,likes,dislikes,comment\_count,comments\_disabled,ratings\_disabled

video\_id : 2kyS6SvSYSE

trending\_date : 17.14.11 (YY.DD.MM)

title : WE WANT TO TALK ABOUT OUR MARRIAGE

channel\_title : CaseyNeistat

category\_id : 22

publish\_time : 2017-11-13T17:13:01.000Z

tags SHANtell martin

views 748374

likes 57527

dislikes 2966

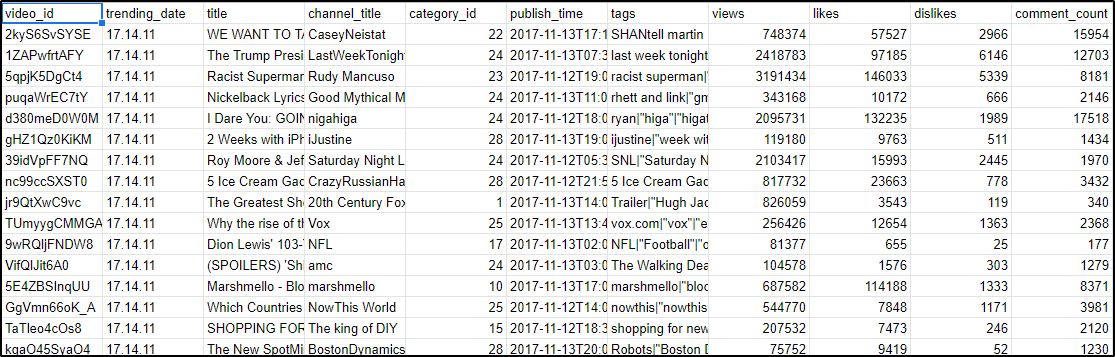
comment\_count 15954

comments\_disabled FALSE

ratings\_disabled FALSE

2kyS6SvSYSE,17.14.11,WE WANT TO TALK ABOUT OUR MARRIAGE,CaseyNeistat,22,2017-11-13T17:13:01.000Z,SHANtell martin,748374,57527,2966,15954,FALSE,FALSE

Data looks as follows.



* category\_title.csv

Data looks as follows.



22,People & Blogs

**Insights on Historical Data**

* **Top 3 videos for which user interaction (views + likes + dislikes +**

**comments) is the highest.**

**Insights on Historical Data**

* **Bottom 3 videos for which user interaction (views + likes + dislikes + comments) is lowest.**
* **Top 3 videos of each category in each year**
  + By number of views
  + By number of comments
  + By number of likes
  + Highest user interaction
* **Top 3 videos in each month**
  + Likes or Dislikes ratio is highest
* **Top 3 videos of each category in each month**
  + By number of views
  + By number of likes
  + By number of dislikes
* **Top 3 channels**
  + By number of total views
  + Likes or Dislikes ratio is highest
  + By number of total comments
* **Top 3 categories**
  + By number of total views
  + Likes or Dislikes ratio is highest
  + By number of total comments
* **Top 3 videos**
  + Value calculated by below formula is highest
    - Views on most recent date / (Recent Date - Publish Date)
* Calculate any 3 videos which got at least 5 comments on every 1000 views
* Calculate any 3 videos which got at least 4 likes on every 100 views
* Number of videos published in each category.

c