INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 6)
Available online at: https://www.ijariit.com

Analysis of user preferred YouTube videos

Divya H.
divyaharids2018@gmail.com
Coimbatore Institute of Technology, Coimbatore, Tamil Nadu

Jane Karuniya J.
karunjaysce.mscds2k18@gmail.com
Coimbatore Institute of Technology, Coimbatore, Tamil Nadu

Srinidhi S.T.
nithishanmugam301@gmail.com
Coimbatore Institute of Technology, Coimbatore, Tamil Nadu

ABSTRACT

In current era of big data, a wide variety of high-volume data having different veracity can be easily collected or generated at a high velocity. Social network data, as well as audio and video in social media and social networking sites, are examples of big data. Embedded in these big data are valuable information and knowledge. Previously unknown and useful information and knowledge from these big data, some big data solutions are in demand. In this project, we explore the big data for detecting various results from YouTube video data so that the user- preferred YouTube viewing can be recognized and the analysis of user- preferred YouTube videos can then be enhanced.

Keywords: *Big Data , You tube, Analysis, Pyspark*

1. INTRODUCTION

Big Data refers to large datasets that could not be analyzed by traditional database systems and processes like RDBMS and existing DataWarehousing systems. Big Data is generally characterized by Huge Volume, High Velocity and High Variety. Companies like Google, YouTube, Facebook, Amazon, Alibaba, Pandora, and Wikipedia are generating and collecting Petabytes of Big data every minute in mutistructured formats likes videos, audios, images, metadata, logs etc. The data generated can be used for Recommendations, formulating business and market strategies using Data Analysis and applying machine learning algorithms. It has been estimated that 2.5 Exabyte of Data is produced every day. The Volume and variety of Big Data makes the task of Data Analysis using existing traditional Data processing techniques extremely challenging. To solve this issue, organizations are shifting towards using multiple servers and using parallel processing to save time and memory. There are different technologies like Hadoop, Spark, HBASE that have been developed and are rapidly evolving to deal with Big Data.

As part of Advanced Databases project, we have extracted and analyzed dataset from YouTube API. The dataset size was 60 MB and consists of key attributes like Video Id, views, likes, Comments and Categories etc. We performed Data Analysis using Apache Spark open source software framework.

YouTube has 1.3 Billion users and 300 hours of video are being uploaded in YouTube every minute. YouTube gets 30 million users every day and nearly 5 Billion views are watched every day. Hence, YouTube has now become very important marketing tool for major companies and entertainment channels. The primary purpose if this project is to find how real YouTube time data can be analyzed to get latest analysis and trends. Along with videos with highest view count, most watched categories, we will find how the user base is increasing and how their interests are changing.

2. DATASET DESCRIPTION

Youtube is the most popular and most used video platform of the world which contains very huge amount of data which is updated constantly. The dataset contains the list of videos that were uploaded in the year 2017 in India. It is an semi-structured dataset with 18736 rows containing attributes like video id, trending date, title, channel title, category id, publishing time, tags, views, likes, dislikes, comment count, thumbnail link, comment disabled, rating disabled, video error/removed and description.

Using all of these attributes will not be necessary, hence the data is preprocessed and then only the necessary attributes are extracted for the analysis that is to be carried out.

3. FRAMEWORKS USED

3.1 Pyspark

Apache spark is an open source clustering computing framework. This has been written in scala language. Pyspark is a collaboration of Apache spark. It can perform both stream processing and batch processing. It is widely used for machine learning and real time streaming analytics.

3.2 Spark SQL

Spark SQL is a component on top of Spark Core that introduces a new data abstraction called SchemaRDD, which provides support for structured and semi-structured data.

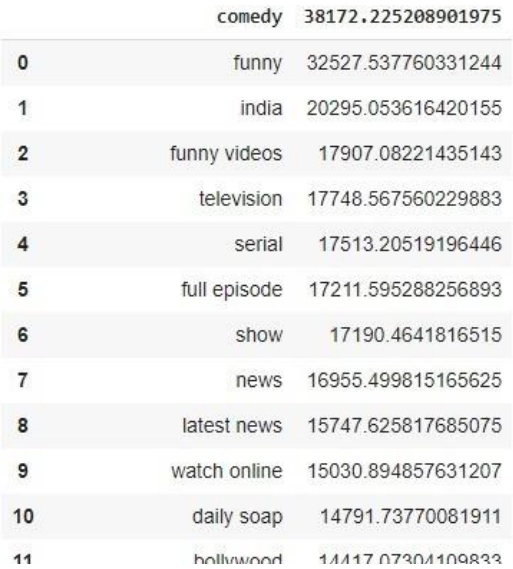
3.3 Queries

Following are the list of queries we tried to answer using Spark.

- Top 20 videos with largest number of views.
- Top 20 videos with largest number of likes.
- Top 20 videos with largest number of dislikes.
- Top 10 viewed categories.
- Most watched categories per year

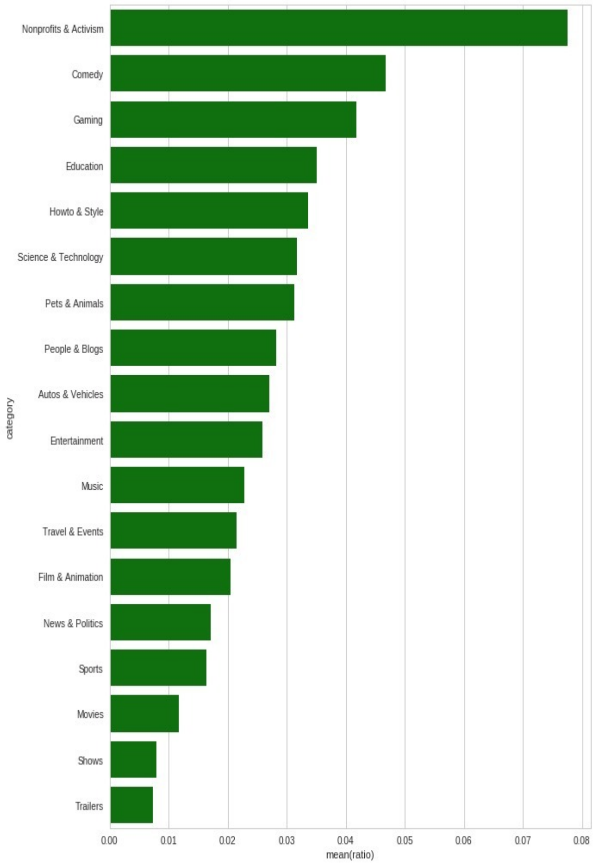
4. RESULTS

4.1 Outputs and Inference

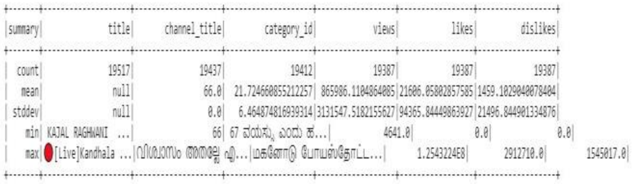


	comedy	38172.225208901975
0	funny	32527.537760331244
1	india	20295.053616420155
2	funny videos	17907.08221435143
3	television	17748.56756229883
4	serial	17513.20519196446
5	full episode	17211.595288256893
6	show	17190.4641816515
7	news	16955.499815165625
8	latest news	15747.625817685075
9	watch online	15030.894857631207
10	daily soap	14791.73770081911
11	bollywood	14417.07304109833

Fig. 1: Shows the list of videos has been sorted from the highest to lowest views in each of the category.

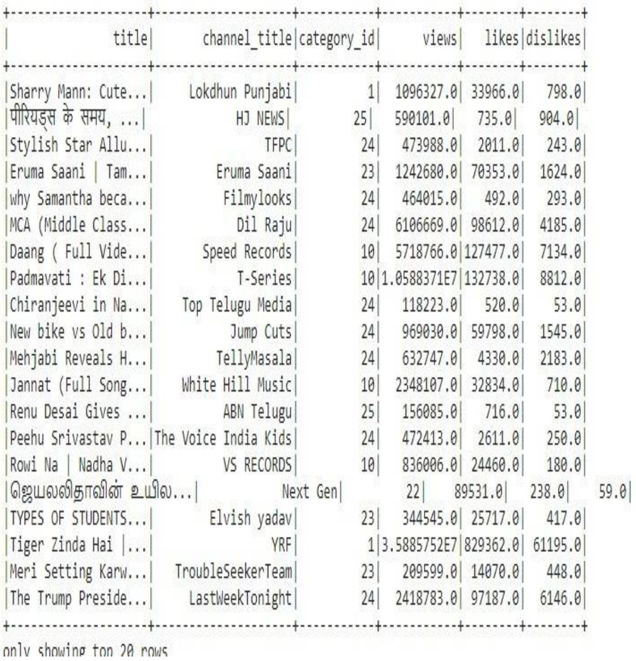


Based on the category each of them has been plotted as a bar graph based on the ratio of views and likes. From this graph we infer that many people have shown interest to watch videos on non profits and activism.



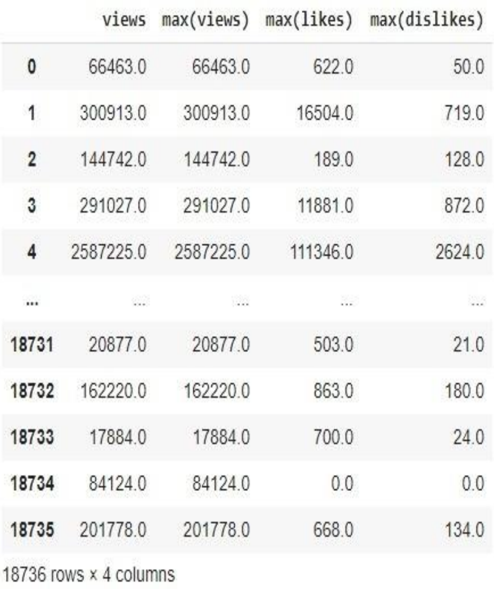
```
+-------+--------------------+--------------------+------------------+------------------+------------------+------------------+
|summary|               title|       channel_title|       category_id|             views|             likes|          dislikes|
+-------+--------------------+--------------------+------------------+------------------+------------------+------------------+
|  count|               19517|               19437|             19412|             19387|             19387|             19387|
|   mean|                null|                66.0| 21.724660855212257| 865986.1104864085|21606.05802857585|1459.1029040078404|
| stddev|                null|                 0.0| 6.464874816939314|3131547.5182155627|94365.84449863927|21496.844901334876|
|    min|KAJAL RAGHWANI   ...|                  66|67 ವಯಸ್ಸು ಎಂದು ಹ...|            4641.0|               0.0|               0.0|
|    max|🔴[Live]Kandhala ...|വിശ്വാസം അതല്ലേ എ...|மகனோடு போயஸ்தோட்ட...|        1.254322E8|         2912710.0|         1545017.0|
+-------+--------------------+--------------------+------------------+------------------+------------------+------------------+
```

This shows the summary of the preprocessed data where only necessary attributes have been chosen. From the summary table displays the videos which has min and max views and the most liked and disliked videos with their channel name and category id.



```
+--------------------+--------------------+-----------+----------+--------+--------+
|               title|       channel_title|category_id|     views|   likes|dislikes|
+--------------------+--------------------+-----------+----------+--------+--------+
|Sharry Mann: Cute...|      Lokdhun Punjabi|          1| 1096327.0| 33966.0|   798.0|
|पीरियड्स के समय, ...|             HJ NEWS|         25|  590101.0|   735.0|   904.0|
|Stylish Star Allu...|                TFPC|         24|  473988.0|  2011.0|   243.0|
|Eruma Saani | Tam...|         Eruma Saani|         23| 1242680.0| 70353.0|  1624.0|
|why Samantha beca...|           Filmylooks|         24|  464015.0|   492.0|   293.0|
|MCA (Middle Class...|            Dil Raju|         24| 6106669.0| 98612.0|  4185.0|
|Daang ( Full Vide...|       Speed Records|         10| 5718766.0|127477.0|  7134.0|
|Padmavati : Ek Di...|            T-Series|         10|1.0588371E7|132738.0|  8812.0|
|Chiranjeevi in Na...|   Top Telugu Media|         24|  118223.0|   520.0|    53.0|
|New bike vs Old b...|           Jump Cuts|         24|  969030.0| 59798.0|  1545.0|
|Mehjabi Reveals H...|          TellyMasala|         24|  632747.0|  4330.0|  2183.0|
|Jannat (Full Song...|    White Hill Music|         10| 2348107.0| 32834.0|   710.0|
|Renu Desai Gives ...|          ABN Telugu|         25|  156085.0|   716.0|    53.0|
|Peehu Srivastav P...|The Voice India Kids|         24|  472413.0|  2611.0|   250.0|
|Rowi Na | Nadha V...|          VS RECORDS|         10|  836006.0| 24460.0|   180.0|
|ஜெயலலிதாவின் உயில...|            Next Gen|         22|   89531.0|   238.0|    59.0|
|TYPES OF STUDENTS...|        Elvish yadav|         23|  344545.0| 25717.0|   417.0|
|Tiger Zinda Hai |...|                 YRF|          1|3.5885752E7|829362.0| 61195.0|
|Meri Setting Karw...|   TroubleSeekerTeam|         23|  209599.0| 14070.0|   448.0|
|The Trump Preside...|     LastWeekTonight|         24| 2418783.0| 97187.0|  6146.0|
+--------------------+--------------------+-----------+----------+--------+--------+
only showing top 20 rows
```

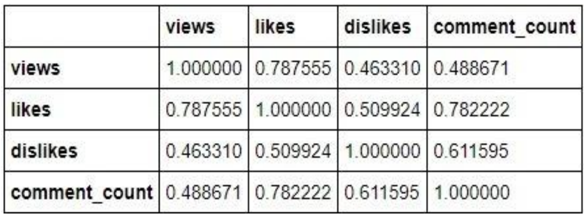
These are list of tables of the dataset after preprocessing which shows the videos and their respective max likes and dislikes with views along with their channel name.



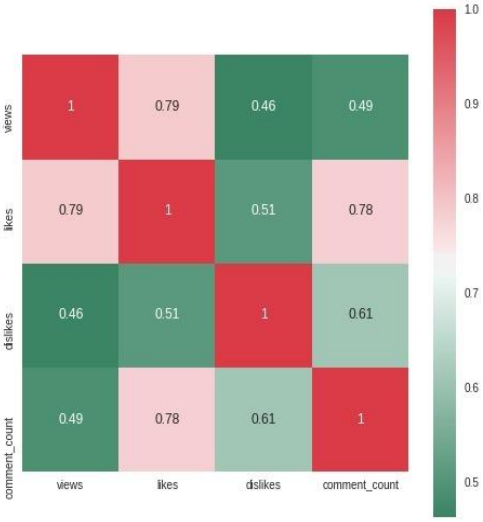
	views	max(views)	max(likes)	max(dislikes)
0	66463.0	66463.0	622.0	50.0
1	300913.0	300913.0	16504.0	719.0
2	144742.0	144742.0	189.0	128.0
3	291027.0	291027.0	11881.0	872.0
4	2587225.0	2587225.0	111346.0	2624.0
...	...	...	...	...
18731	20877.0	20877.0	503.0	21.0
18732	162220.0	162220.0	863.0	180.0
18733	17884.0	17884.0	700.0	24.0
18734	84124.0	84124.0	0.0	0.0
18735	201778.0	201778.0	668.0	134.0

18736 rows × 4 columns

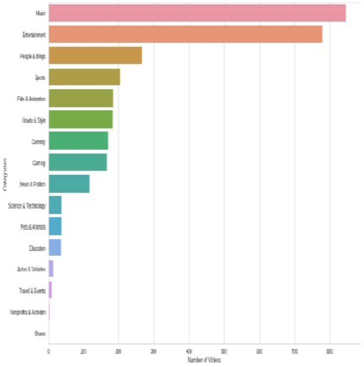
For every video in the dataset the number of views, likes and dislikes has been listed out. From this we can see that even the most liked or viewed video also has been disliked and the video with most dislikes has also received some likes.



	views	likes	dislikes	comment_count
views	1.000000	0.787555	0.463310	0.488671
likes	0.787555	1.000000	0.509924	0.782222
dislikes	0.463310	0.509924	1.000000	0.611595
comment_count	0.488671	0.782222	0.611595	1.000000



The correlation matrix tells us the movement of one variable with another variable. The correlation lies between -1 to +1. From the above correlation matrix we can see that all the values are positive and hence we can tell it is a positive correlation. This matrix infers that how each of the attributre correlates with another attribute. It has been represented in a heat map for better visualization.



From the final table(figure 8) we infer that many of the people preferred to watch music videos and finally we came to conclude that many people prefer to watch music videos.

5. CONCLUSION

The visualizations has been carried out using apache spark and spark SQL. The Spark Sql has been used for implementing queries to get the most liked , views and disliked video and the video count in each category. The videos in category has also been displayed. From the visualizations we can conclude that music category has gained more views and likes. People are more interested in music videos and contents. Hence this analysis and visualization will be helpful for the youtubers to decide what are people’s interest and what type of videos can be uploaded to gain more attention. Also which factor attracts people towards the video.

6. REFERENCES

[1] CHENG Xu et.al, “Understanding the YouTube and their data”, [IEEE,2014]

[2] M. Dehghani, M. K. Niaki, I. Ramezani, and R. Sali,“Evaluating the influence of youtube advertising for attraction of young customers,” Computers in Human Behavior, vol. 59, pp. 165 – 172, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563216300450