Analysing YouTube Data Using K-Means Clustering

G. Ravali[1], P.Swathi Moulika[2], A.A parna[3], J.Abhinaya[4], T.Anuradha [5]

[1,2,3,4] IV/IV B.Tech ,[5]Professor

VELAGAPUDI RAMAKRISHNA SIDDHARTHA ENGINEERING COLLEGE
Dept. of Information Technology ,kanuru,A.P,India

Abstract

YouTube is one of the most popular social networking sites which have millions of users posting and viewing different kinds of videos. Having different channels posting different categories of videos regularly and users from different parts of the globe giving feedback on these videos, YouTube is one of the best source of big data. Previous analysis on YouTube data are dealing with users' sentimental analysis related to different issues. This paper mainly concentrates on analysing YouTube data of a taken time period in terms of different parameters like category, channel and location. It deals with analysing YouTube data based on data mining clustering algorithm named K-Means. As the existing relational databases are lagging behind handling voluminous data, people started moving towards NOSQL databases. One such popular NOSQL database is MongoDB which has high scalability. Results are analysed visually by using tools like Weka, Rapidminer and Tableau.

Keywords — YouTube, big data mining, k means clustering, MongoDB , NOSQL,JSON

1.INTRODUCTION

YouTube is one of the most popular video sharing websites which is used by many users worldwide[1]. There are different channels posting different kinds of videos on You Tube related to different categories frequently. Videos that are posted on this site belong to various categories like education , comic ,music ,entertainment ,lifestyle ,fashion etc. The videos that are present in this website can be rated based on users' likes, dislikes or comments on the videos posted[2].Most of the existing studies on YouTube is based on like-dislike count related to healthcare issues and sentimental analysis based on the positive or negative comments given for a video. This paper concentrates on various video categories present in YouTube and analyzing those categories based on different parameters like Like Count, Comment Count , Dislike Count, Channel Title, region etc. Few categories are chosen for analysis over a time period ranging from October 2016 to January 2017.In this time period the YouTube videos data is collected and stored in a database named MongoDB [3] which is a NOSQL database, which has the capability to store larger amount of data compared to SQL databases. On the entire data collected from YouTube API V3[4] over a taken time period , K-Means algorithm is applied for clustering the data. Once the clusters are generated they are going to provide analysis reports about the category to which ? majority of videos belong , the channel to which the videos belongs and the region wise like count of videos. All these reports are visually represented using few famous data mining tools like Weka, RapidMiner and Tableau.

2.TECHNIQUES USED

2.1MONGODB:

Relational databases have been used from decades for the purpose of storing data at the back end for business applications or web applications. Scalability is one of the key issue faced by many data stores. Relational databases have less capability for horizontal scaling because of which NOSQL databases were developed to meet the scaling demands. The most important feature of NoSQL databases is "Share Nothing" horizontal scaling[5]which provides replicating the data over many servers because of which large number of read and write operations can be handled in NoSQL.

MongoDB is one such NoSQL database which is used in this paper. It is a cross-Platform, document oriented database which has documents in JSON format. It provides high performance, high availability and easy scalability. One of the key concept of MongoDB is sharding[6]. It is a method of distributing data across different machines to support deployment of large data sets which is done by horizontal scaling.

2.2K-Means Algorithm:

K-Means algorithm is one of the famous centroid based clustering algorithm which is used to partition the dataset automatically into k groups[7]. It starts by selecting k initial cluster centers (centroids) and then iteratively refining them until the algorithm is converged i.e. each data point remains in the same cluster even after the next iteration. The entire algorithm runs on a parameter named as Euclidean Distance, which helps in calculating the distance between the data points using the formula as shown :

$$F=\sum^{k}_{j=1} \sum^{n}_{i=1} (x_i - c_j)^2$$

The steps involved in the execution of the algorithm are as follows:

1)Let 'D' be the data set which consists of a set of data points {x1,x2,x3…..xn}.
2) Let C={c1,c2….ck} be the set of cluster centers (centroids).
3) Compute the distance between each data point to the cluster centers.
4) Assign the data point to the cluster center whose distance from the cluster center is minimum when compared to other cluster centers.
5) Recalculate the new cluster center C_J by calculating the average of the data points x_i in the respective cluster.
6) Repeat the steps (4) and (5) until convergence and return the final cluster centers (centroids) {c1,c2….ck} along with the data points in their respective clusters.

3.LITERATURE SURVEY:

Jennifer Keelan , Vera Pavri-Garcia had published a paper named "YouTube as a source of information on immunization" which deals with collecting all unique videos basing on two keywords named 'vaccination ' and 'immunization'. All the videos with English language content containing information about human immunization are collected [8]. User's interaction with the videos is measured using viewcount and the viewer's reviews. Basing on the reviews the videos are categorized into negative if the message in the videos is portraying immunization negatively, positively if the video speaks in favour of immunization and ambiguous if there is no specific information provided. This is achieved by using technique called Weighted K-statistics (Cohen's Kappa) and the results obtained shows that 49 videos (32%) are in the negative category , 73(48%) videos are in the positive category and 31 (20%) are in the ambiguous category.
Heather Molyneaux , Susan O'Donnell , Kerri Gibson and Janice Singer had published a paper named "Exploring the Gender Divide on YouTube:An Analysis of the creation and reception of Vlogs". This paper deals with the way men and women interact with Vlogs and the way they react to Vlogs[9]. Here Dual analytical approach is used to analyse the production and reception of Vlogs by the people. Here data related to 15 days is taken as input during this time period the Vlogs of length more than 3min and non English vlog are eliminated and the remaining vlogs are considered for analysis. On analysing the results it was found that there is a gender imbalance in both the reception and creation of vlogs. The results tell that women are posting less frequently than men and women vloggers are more likely to interact with other vloggers compared to men.

Hye-jin peak , Kyongseok kim and Thomas Hove had published a paper named "content analysis of antismoking videos on the YouTube" .This paper deals with the promotional of antismoking videos among the YouTube. The main technique in this process is Box's M test. In order to analyze the data they used three methods like Message sensitive value, Message appeals and Relationships with viewer's response[10]. In message appeal the analysis is done by dividing the comments into three categories like Threat appeal, Social appeal and Humor appeal .The final result tells that antismoking videos are more popular and widespread among the world. When the keyword antismoking is used only a few videos are retrieved but using the keyword Cigarette more number of videos are retrieved.

Ahmad Ammari , Vania Dimitrova and Dimoklis Despotakis had published a paper named "Semantically Enriched Machine Learning Approach to Filter YouTube Comments for Socially Augmented User Models". This paper deals about the analysis of the noisy data in the social media like YouTube , Flickr and delicious and is used to detect the noisy comments by using a technique called Roadmap which is a pattern mining technique[11]. These sites are being mostly used in the present world for exchanging the information with others. In YouTube the comments will be posted for a particular video, some of the comments will be positive or negative related to that video and some comments are totally irrelevant .These irrelevant are termed as noisy comments. The output results shows the comparison of the comments for a particular video before and after filtering the noisy comments.

4.PROPOSED METHOD:

The main aim of the paper is to analyze the YouTube data which is gathered from the YouTube developer API V3. The data obtained from API is in JSON (JavaScript Object Notation) format which is unstructured and is converted to structured format by using JSON parsers and Google spreadsheets method. The structured data is dumped into Mongo DB which is used for storing and retrieving the data. Preprocessing steps like filling the missing values by using '0' and 'NULL' are applied on the collected information. Now on this information, K-Means clustering algorithm is applied and clusters are obtained. The obtained results are imported into data mining tools named Weka, RapidMiner and Tabluea. These tools help in providing the visual representation[12] of the clusters in different formats. The generated analysis reports provides us information about the number of data points in each cluster ,the number of videos belonging to each category, the category to which maximum videos belong ,the number of videos belonging to a particular channel and the like count of videos basing on regions. In order to get further in depth details about clusters formed we can once again bisect each cluster that is formed iteratively using k means and generate more details about the clusters basing on various parameters. Fig 1 shows the sequence of steps used in the experimentation.

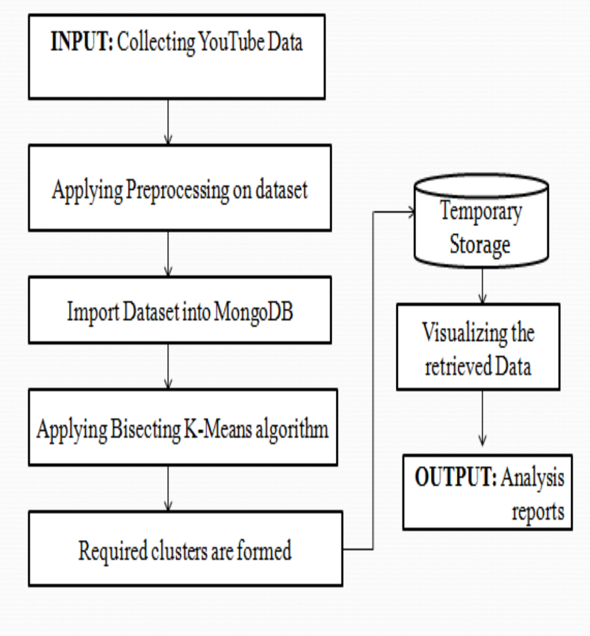
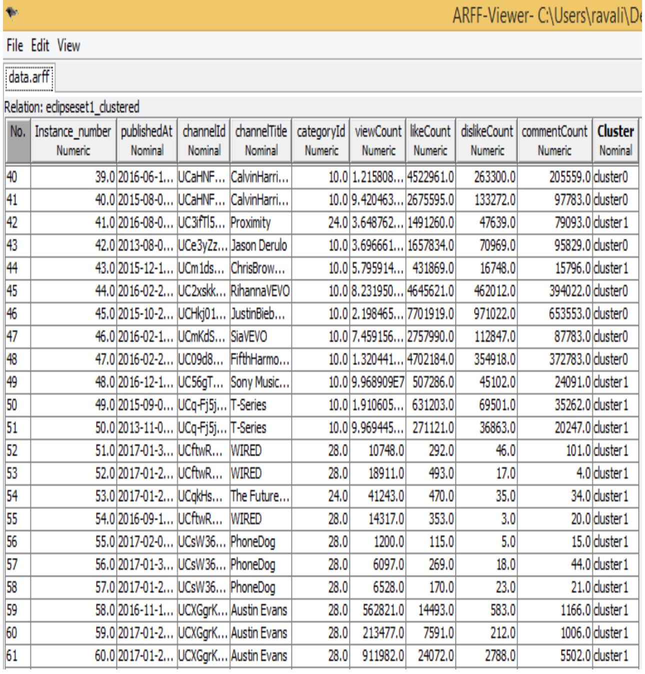

INPUT: Collecting YouTube Data
Applying Preprocessing on dataset
Import Dataset into MongoDB
Applying Bisecting K-Means algorithm
Required clusters are formed
Temporary Storage
Visualizing the retrieved Data
OUTPUT: Analysis reports

Fig 1: Project Flow

5.EXPERIMENTAL RESULTS

Analysis results are obtained basing on the attributes of the dataset. There are two clusters formed after performing K means clustering. The Fig 2 represents the cluster along with its data points that are assigned to that cluster. Attributes like- like count, view count etc. which are having similar values come into the same cluster either cluster '0' or cluster '1'. Fig 3 represents the number of data points in each cluster. Majority of the data points come under cluster 1.



Relation: eclipseset1_clustered

No.	Instance_number Numeric	publishedAt Nominal	channelId Nominal	channelTitle Nominal	categoryId Numeric	viewCount Numeric	likeCount Numeric	dislikeCount Numeric	commentCount Numeric	Cluster Nominal
40	39.0	2016-06-1...	UCaHNF...	CalvinHarri...	10.0	1.215808...	4522961.0	263300.0	205559.0	cluster0
41	40.0	2015-08-0...	UCaHNF...	CalvinHarri...	10.0	9.420463...	2675595.0	133272.0	97783.0	cluster0
42	41.0	2016-08-0...	UC3ifTl5...	Proximity	24.0	3.648762...	1491260.0	47639.0	79093.0	cluster1
43	42.0	2013-08-0...	UCe3yZz...	Jason Derulo	10.0	3.696661...	1657834.0	70969.0	95829.0	cluster0
44	43.0	2015-12-1...	UCm1ds...	ChrisBrow...	10.0	5.795914...	431869.0	16748.0	15796.0	cluster1
45	44.0	2016-02-2...	UC2xskk...	RihannaVEVO	10.0	8.231950...	4645621.0	462012.0	394022.0	cluster0
46	45.0	2015-10-2...	UCHkj01...	JustinBieb...	10.0	2.198465...	7701919.0	971022.0	653553.0	cluster0
47	46.0	2016-02-1...	UCmKdS...	SiaVEVO	10.0	7.459156...	2757990.0	112847.0	87783.0	cluster0
48	47.0	2016-02-2...	UC09d8...	FifthHarmo...	10.0	1.320441...	4702184.0	354918.0	372783.0	cluster0
49	48.0	2016-12-1...	UC56gT...	Sony Music...	10.0	9.968909E7	507286.0	45102.0	24091.0	cluster1
50	49.0	2015-09-0...	UCq-Fj5j...	T-Series	10.0	1.910605...	631203.0	69501.0	35262.0	cluster1
51	50.0	2013-11-0...	UCq-Fj5j...	T-Series	10.0	9.969445...	271121.0	36863.0	20247.0	cluster1
52	51.0	2017-01-3...	UCftwR...	WIRED	28.0	10748.0	292.0	46.0	101.0	cluster1
53	52.0	2017-01-2...	UCftwR...	WIRED	28.0	18911.0	493.0	17.0	4.0	cluster1
54	53.0	2017-01-2...	UCqkHs...	The Future...	24.0	41243.0	470.0	35.0	34.0	cluster1
55	54.0	2016-09-1...	UCftwR...	WIRED	28.0	14317.0	353.0	3.0	20.0	cluster1
56	55.0	2017-02-0...	UCsW36...	PhoneDog	28.0	1200.0	115.0	5.0	15.0	cluster1
57	56.0	2017-01-3...	UCsW36...	PhoneDog	28.0	6097.0	269.0	18.0	44.0	cluster1
58	57.0	2017-01-2...	UCsW36...	PhoneDog	28.0	6528.0	170.0	23.0	21.0	cluster1
59	58.0	2016-11-1...	UCXGgrK...	Austin Evans	28.0	562821.0	14493.0	583.0	1166.0	cluster1
60	59.0	2017-01-2...	UCXGgrK...	Austin Evans	28.0	213477.0	7591.0	212.0	1006.0	cluster1
61	60.0	2017-01-2...	UCXGgrK...	Austin Evans	28.0	911982.0	24072.0	2788.0	5502.0	cluster1

Fig 2 : Instances with cluster number as shown in weka(Majority cluster 1)

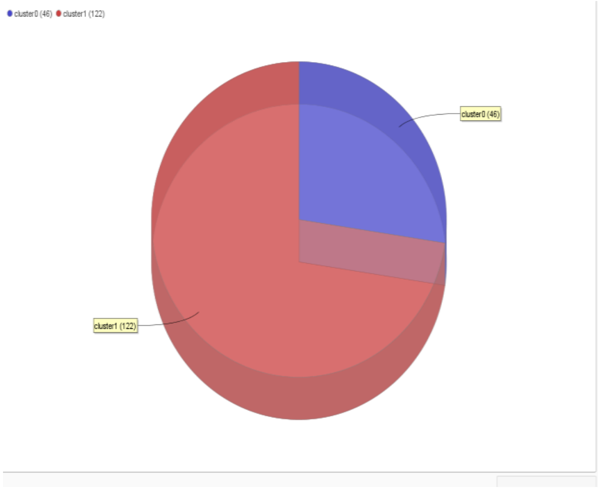
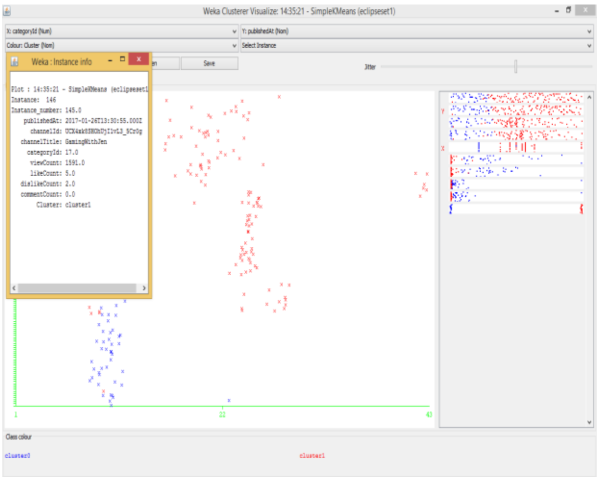


Fig 3: Data points in Cluster as shown in RapidMiner

The clusters formed are pictorially represented in Fig 4. The data points which are highly concentrated in an area represent more similarity. On clicking on any data point its details are displayed in a new window named Instance info as show in Fig 4.



```
Plot : 14:35:21 - SimpleKMeans (eclipseset1
Instance:  146
Instance_number: 145.0
    publishedAt: 2017-01-26T13:30:55.000Z
      channelId: UCX4ak8SHOhDjIlvL3_SCr0g
   channelTitle: GamingWithJen
     categoryId: 17.0
      viewCount: 1591.0
      likeCount: 5.0
   dislikeCount: 2.0
   commentCount: 0.0
        Cluster: cluster1
```

Fig 4: Instance info as shown in weka

As per YouTube norms each category is given a specific category id and some of the categories are shown in Fig 5. Fig 6 represent the category to which majority of videos belong during the taken time period from October 2016 to January 2017.

Category id	Category name
2	Autos & Vehicles
10	Music
17	Sports
18	Short Movies
19	Travel & Events
20	Gaming

22	**People & Blogs**
25	**News & Politics**
27	**Education**
42	**Shorts**

Fig 5: Category Details

Fig 6 clearly represent that category id ‘10’ named Music has the maximun number of videos in the dataset followed by category id ’25’ named news and politics , category id ‘20’ named Gaming.The minimum number of videos are related to category id ‘2’ named Autos and Vehicles, category id ‘19’ named Travel & Events.The remaning categories like sports,Education etc falls in other frequency range.

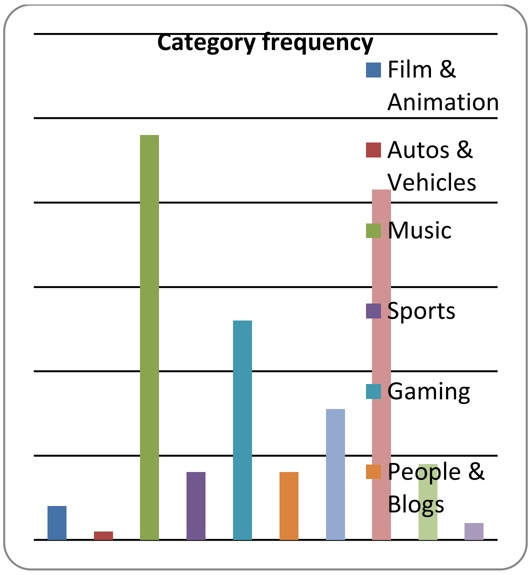


Fig 6: Category Frequency

Basing on the ChannelTitle , a pie chart is formed which describes the number of videos belonging to each ChannelTitle. From Fig 7 majority of the videos belongs to Lahari music T-series, followed by Typical gamer channel and One direction VEVO channel and least number of videos belong to channel Christina Perri channel.

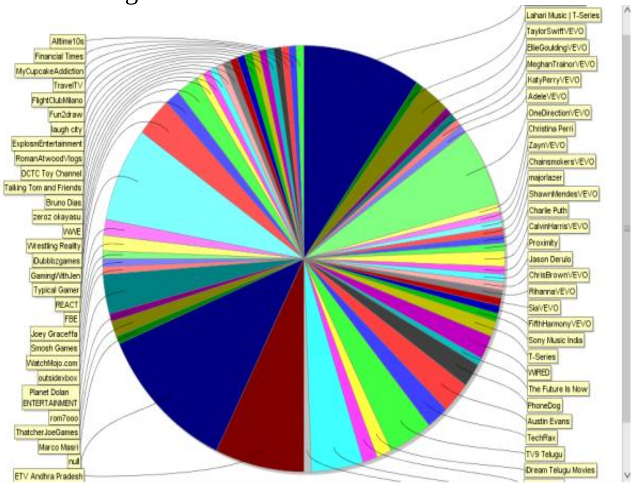
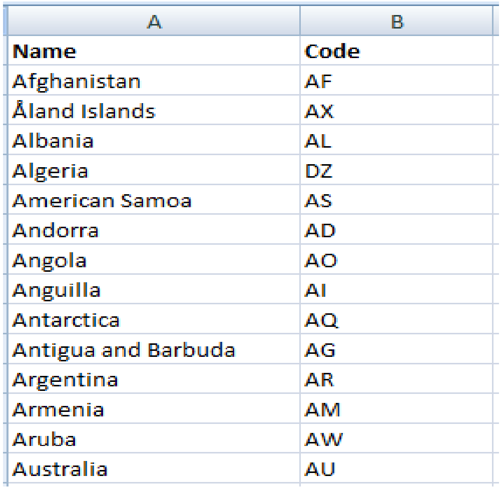


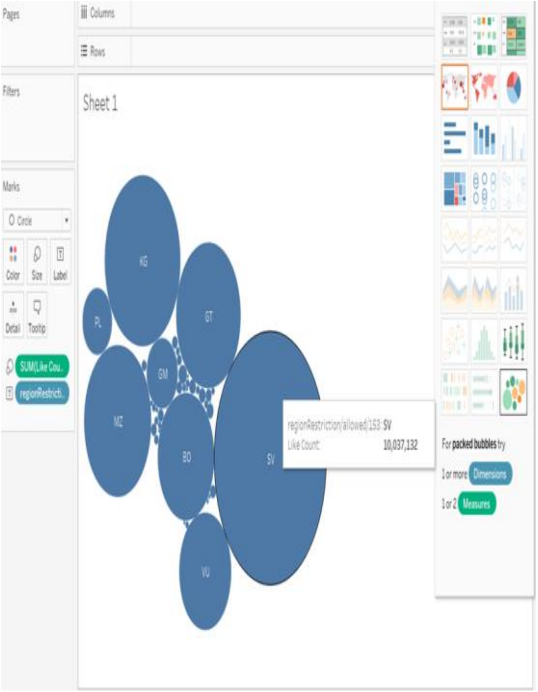
Fig 7: ChannelTitle as shown in RapidMiner

YouTube videos are played in different countries/regions.The likecount of videos varies from place to place.Fig 9 deals with a single category named music (Categoy id ’10’). The likecount of music related videos is maximum in country SV(El Salvador) followed by regions like KG(Kyrgyzstan),MZ(Mozambique), GT(Guatemala), BO(Bolivia),etc .



A	B
Name	**Code**
Afghanistan	AF
Åland Islands	AX
Albania	AL
Algeria	DZ
American Samoa	AS
Andorra	AD
Angola	AO
Anguilla	AI
Antarctica	AQ
Antigua and Barbuda	AG
Argentina	AR
Armenia	AM
Aruba	AW
Australia	AU

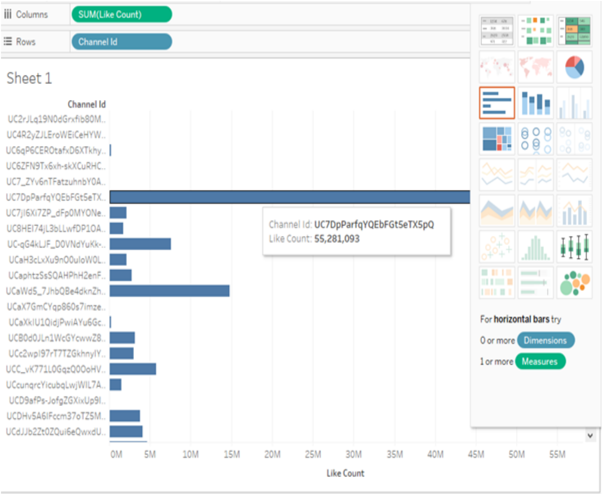
Fig 8 : Region codes



regionRestriction/allowed/153: SV
Like Count: 10,037,132

Fig 9 : Location wise Like count of videos in music category as shown in Tableau

A particular category consists of a number of different channels in YouTube.Basing on the channelid the likecount of videos are compared in Fig 10.The channelid ‘UC7DpParfqYQEbFGt5eTX5pQ”
has the maximum likecount followed by channelid ‘UCaWd5_7JhbQBe4dknZhsHJg’.



Columns: SUM(Like Count)

Rows: Channel Id

Sheet 1

Channel Id: UC7DpParfqYQEbFGt5eTX5pQ
Like Count: 55,281,093

Fig 10: Likecount based on Channelid as shown in Tableau.

6.CONCLUSION

YouTube data analysis based on the taken dataset was done with respect to the category of videos, number of videos belonging to a specific channel and the like count of videos region wise. The results proved that majority of videos posted on You Tube for the taken time period belonged to music category and least number of videos were of Autos and Vehicles category. With in the music category, channel titled Lahari music T-series has posted maximum number of videos and the maximum like count is achieved from the regions SV,KG,MZ,GT.Basing on channelid the likecount is maximum for channel ‘C7DpParfqYQEbFGt5eTX5pQ’.

REFERENCES:

[1] YouTube, L. L. C. "YouTube." Retrieved 27 (2011): 2011.

[2] Burgess, Jean, and Joshua Green. YouTube: Online video and participatory culture. John Wiley & Sons, 2013.

[3] Banker, Kyle. MongoDB in action. Manning Publications Co., 2011.

[4] Reuter, Christian, and Simon Scholl. "Technical Limitations for Designing Applications for Social Media." Mensch & Computer Workshopband. 2014.

[5] Arora, Rupali, and Rinkle Rani Aggarwal. "Modeling and querying data in mongodb." International Journal of Scientific and Engineering Research 4.7 (2013).

[6] Chodorow, Kristina. Scaling MongoDB: Sharding, Cluster Setup, and Administration. " O'Reilly Media, Inc.", 2011.

[7] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.

[8] Keelan, Jennifer, et al. "YouTube as a source of information on immunization: a content analysis." jama 298.21 (2007): 2482-2484.

[9] Molyneaux, Heather, et al. "Exploring the gender divide on YouTube: An analysis of the creation and reception of vlogs." American Communication Journal 10.2 (2008): 1-14.

[10] Paek, Hye-Jin, Kyongseok Kim, and Thomas Hove. "Content analysis of antismoking videos on YouTube: message sensation value, message appeals, and their relationships with viewer responses." Health education research 25.6 (2010): 1085-1099.

[11] Ammari, Ahmad, Vania Dimitrova, and Dimoklis Despotakis. "Semantically enriched machine learning approach to filter YouTube comments for socially augmented user models." UMAP (2011): 71-85.

[12] Soukup, Tom, and Ian Davidson. Visual data mining: Techniques and tools for data visualization and mining. John Wiley & Sons, 2002.