Trending YouTube Video Statistics

April 18, 2019

Team members:
Elmaddin Karimov
Batbileg Enkhbat
Mukta Jathar

Table of Contents

1. INTRODUCTION

Social media play an increasingly more important role in the life of society as users spend tons of time online and they view social media as an important source of information about topics which users are concerned with. YouTube, being one of the mainstream social media, allows users to share their videos online, create their channels and, thus, create their virtual communities united by common interests.

In terms of content creation, every minute of the day 300 hours of video are uploaded onto YouTube platform where the popularity trend changes just as fast. On the other side, YouTube gets over 30 million visitors a day who are eager to consume the content shared by the creators.

YouTube also becomes a tool to promote products and services for entrepreneurs and enterprises. YouTube allows watching diverse videos and channels which may vary in their content. However, the popularity of videos determines their availability to the audience that influences the perception of information by the audience. For instance, the video that has the highest number of views becomes more and more popular and becomes mainstream, even though in its essence the video may have little cultural value or poor messages. In fact, the point is to make videos performable.

What makes YouTube popular and mainstream social media is its performativity because the audience receive what it wants and expects from the media, the performance, the show and entertainment with the possibility of the further communication and even interaction with other users of YouTube. Often the quantity of views determines the popularity of videos in YouTube and users mistakenly associate the quality of the video with the number of views.

2.1 EXTRACTION & TRANSFORMATION:

Data source:

This dataset is obtained from Kaggle.com and is a daily record of up to 200 top trending YouTube videos for various countries/regions. Data is available for USA, Great Britain, Germany, Canada, and France among other regions.

Each region has its own csv and json file. The csv file includes various columns like video title, channel name, publish time, tags, views count, number of likes and dislikes, comment count,etc.

The csv data also includes a category_id field which varies between regions. To retrieve the categories (E.g. Travel & Events, Autos, Education, etc.) associated with each category_id for a specific video, we have to find it in the associated JSON.

Transformation:

A clean dataset was created containing data for three regions (US, Canada, Great Britain) by combining data from three csv and three JSON files. Jupyter Notebook was used to achieve this.
Steps:

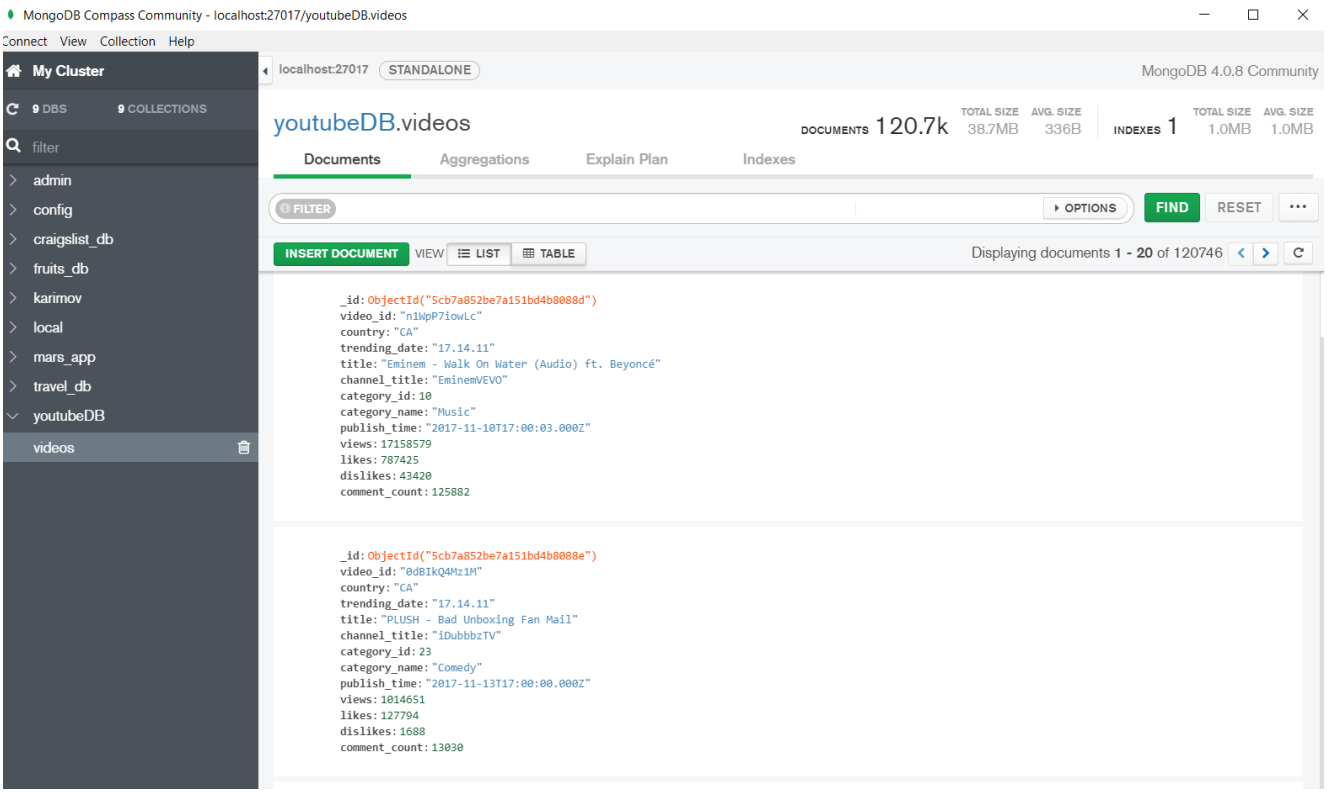
- Read each of the three csv files to a pandas data frame.
- Cleaned the data frames to include only relevant columns that could be useful for analysis.
- The csv data source did not include any column to indicate which country the data belongs to. Hence, added a new column ‘Country’ to the data frames.
- Category name associated with each category Id was retrieved and mapped from JSON files for each region. The JSON file was nested, Python ‘For’ loop was used to read and store the data for ‘category id’, ‘country’ and ‘category name’ from the JSON. The category name was then mapped and added as a new column into the data frames created for each of the regions.
- The data frames for the three regions were then combined into one single data frame containing 120k+ rows.
- The clean data file was exported to csv for analysis and also loaded to a database.

Screenshot of final data-frame:

ID	video_id	country	trending_date	title	channel_title	category_id	category_name	publish_time	views	likes	dislikes	comment_count
1	n1WpP7iowLc	CA	17.14.11	Eminem - Walk On Water (Audio) ft. Beyoncé	EminemVEVO	10	Music	2017-11-10T17:00:03.000Z	17158579	787425	43420	125882
2	0dBIkQ4Mz1M	CA	17.14.11	PLUSH - Bad Unboxing Fan Mail	iDubbbzTV	23	Comedy	2017-11-13T17:00:00.000Z	1014651	127794	1688	13030
3	5qpjK5DgCt4	CA	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	Comedy	2017-11-12T19:05:24.000Z	3191434	146035	5339	8181
4	d380meD0W0M	CA	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	Entertainment	2017-11-12T18:01:41.000Z	2095828	132239	1989	17518
5	2Vv-BfVoq4g	CA	17.14.11	Ed Sheeran - Perfect (Official Music Video)	Ed Sheeran	10	Music	2017-11-09T11:04:14.000Z	33523622	1634130	21082	85067

2.2 LOADING OF DATA TO DATABASE:

MongoDB was used to load the extracted and transformed data. Since the dataset has been derived by stitching together data for different regions, in case a need arises in future to accommodate different types/structure of data for different regions, MongoDB will allow that.



4 CONCLUSIONS

The dataset has been extracted, transformed and loaded. YouTube content creators could utilize the database to perform analysis on what makes a certain video popular more than others. On the other hand, enterprises could utilize the database to obtain marketing insights and gain advantage over the competitors.

DATA SOURCE:

This dataset is available on Kaggle.com and is a daily record of the top trending YouTube videos.

https://www.kaggle.com/datasnaek/youtube-new#CA_category_id.json