Project 1-README

April 27, 2018

0.1 Abstract-

For this project, Movie domain is chosen and real time data from Twitter and Youtube Social media sites are extracted by using web scraping method. Here, I have scrapped IMDB page of Twitter website to model the database.

0.2 Files in Project 1 folder-

0.3 IMDB_tweet_Extractor.ipynb-

This file contains the code used for extracting real time rating data from IMDB webpage of twitter. Data is extracted using Twitter API. All the movie ratings along with user tweet id and timpstamp when tweet was posted is extracted in Tweet_ratings.csv file

0.4 backbone.py-

Backup code for IMDB_tweet_Extractor

0.5 Twitter_Wordcloud_most_used_hashtags.ipynb-

This file fetches streaming tweets related to a topic you speficy. For this project, I have made use of keywords like:these are most recently released movies-

black panther','ready player one','pacific rim','justiceleague','coco','gringo','rampage' for collecting tweets

This file streams real time tweets for these specific words and forms wordcloud for most used hash tagged words.

0.6 Project One.ipynb

Explains intricate details of how entire database is formed.

0.7 Brief description of the tables used in databse-

-Movies- contains data related to movie name, genres and year of release

-Ratings dataset- This data is picked from existing Grouplens database has has movie ratings from 0-5 scale

-Tag- contains all the real time hashtags extracted from IMDB page of twitter

-Tweet_ratings- conatains all the ratings of movies posted on IMDB page of twitter

-Youtube- contains raw data from IMDB channel of Twitter

0.8 Question answered in database

For answering this question, I am using Ratings from table ratings i.e, from Grouplens dataset as well as real time rating data collected from IMDB page of twitter

0.9 What is the best time to post?

Query 1 on ratings table of grouplens dataset

0.9.1 select count(*) as maximum_posts,timestamp from ratings_hourly group by timestamp order by timestamp DESC

0.9.2 Query result

By this query, we can say since at 23rd hour we get the maximum posts so 11 pm is the best time to post

Query 2 on ratings table of real time twitter dataset collected from IMDB page

0.9.3 select count(*) as tweet_count, Tweet_timestamp from tweet_ratings_hourly group by Tweet_timestamp order by tweet_count DESC

0.9.4 Query result

By this query, we can say since at 21st hour we get the maximum tweets so 9 pm is the best time to post