INFO 6210 Movie Database

YURU LI
ZHIHUI XIN

Abstract

The domain of our project is movie. Our goal is to use the data we get from the internet (social media, web scraper, API, raw data, twitter, Instagram, etc.) to build a database which include movie information, actor/actress Instagram data, movie recommend twitter account's tweet data, Oscar winner's data, etc. Also, we are going to create a newsfeed as an interface to query this data.

Keywords
Movie, database, social media, movie information

I. Introduction

When user input an old movie name or an actor name, we do some algorithm by connecting the data tables, then returns an in-theater movie that the user may want to see, at the same time show some related comments or recent news.

People now really love watching movie on internet. Some people have specific aim of what to watch, other people have no idea and they really want to watch movies to relax.

They might have favorite actors/actress, but they don't really know what movies they play, and sometimes people just want to watch special genre like action/comedy. So we build the database of movies, and people could search by genre/actors/title. The database will help people find some related movies. It helps people to make a better choice.

II. Data Resource

1. Old movie information from Kaggle

We got past ten years movie information from Kaggle, but it has too much useless information, and the format need to be cleaned.

old_movie_id	movie_title	director_name	main_actor	genre_id
1	On the Road	Walter Salles	Kristen Stewart	3
3	Enchanted	Kevin Lima	Jeff Bennett	4
4	Dancer, Texas Pop. 81	Tim McCanlies	Ethan Embry	6
5	TMNT	Kevin Munroe	Chris Evans	3
6	Sunshine	Danny Boyle	Chris Evans	7
7	Videodrome	David Cronenberg	Debbie Harry	12
8	Alvin and the Chipmunks: The Squeakquel	Betty Thomas	Amy Poehler	17
9	UnDivided	Sam Martin	Sam Adams	18
10	Universal Soldier: The Return	Mic Rodgers	Michael Jai White	7
11	Panic Room	David Fincher	Kristen Stewart	8
12	Taken 3	Olivier Megaton	Liam Neeson	8
13	Sideways	Alexander Payne	Virginia Madsen	3
14	The Game of Their Lives	David Anspaugh	Gerard Butler	1
15	Poultrygeist: Night of the Chicken Dead	Lloyd Kaufman	John Karyus	6
16	Antitrust	Peter Howitt	Tyler Labine	2

2. New movie information from IMDB

From IMDB, we use web scraper to get the information of latest movies. And we add real-time updated function.

Result Grid Filter Rows: Search Edit: Export/Import:

new_movie_id	title	runtime	genre	director	stars	outline
0	Penguins	76 min	Documentary	Alastair Fothergill|Jeff Wilson	Ed Helms	The story of Steve; an Adé
1	Breakthrough	116 min	Biography|Drama	Roxann Dawson	Chrissy Metz|Topher Grace|Josh Lucas|Marcel…	When her 14-year-old son
2	Hail Satan?	95 min	Documentary	Penny Lane	Jex Blackmore|Nicholas Crowe|Lucien Greavesl…	A look at the quick rise and
3	Little Woods	105 min	Crime|Drama|Western	Nia DaCosta	Lily James|Tessa Thompson|Luke Kirby|Lance…	A modern Western about t
4	High on the Hog	85 min	Action|Crime|Drama|Thriller	Tony Wash	Sid Haig|Joe Estevez|Robert Z'Dar|Fiona Dome…	With a potent strain of pot
5	Pet Sematary	101 min	Horror|Mystery|Thriller	Kevin Kölsch|Dennis Widmyer	Jason Clarke|Amy Seimetz|John Lithgow|Jeté L…	Dr. Louis Creed and his wi
6	Us	116 min	Horror|Thriller	Jordan Peele	Lupita Nyong'o|Winston Duke|Elisabeth Moss|Ti…	A family's serene beach va
7	The Best of Enemies	133 min	Biography|Drama|History	Robin Bissell	Taraji P. Henson|Sam Rockwell|Babou Ceesay|…	Civil rights activist Ann Atw
8	Unplanned	106 min	Drama	Chuck Konzelman|Cary Solomon	Ashley Bratcher|Brooks Ryan|Robia Scott|Jared…	Abby Johnson is one of the
9	How to Train Your Dragon: The Hidden World	104 min	Animation|Action|Adventure|Comedy|Family|Fa…	Dean DeBlois	Jay Baruchel|America Ferrera|F. Murray Abraha…	When Hiccup discovers To
10	The Curse of La Llorona	93 min	Horror|Mystery|Thriller	Michael Chaves	Linda Cardellini|Raymond Cruz|Patricia Velasqu…	Ignoring the eerie warning
11	Kalank	166 min	Drama|Romance	Abhishek Varman	Varun Dhawan|Alia Bhatt|Aditya Roy Kapoor|Ma…	Drama set in the 1940s du
12	Under the Silver Lake	139 min	Comedy|Crime|Drama|Mystery|Thriller	David Robert Mitchell	Andrew Garfield|Riley Keough|Topher Grace|Ca…	Sam; intelligent but withou
13	Rafiki	83 min	Drama|Romance	Wanuri Kahiu	Samantha Mugatsia|Neville Misati|Nice Githinji|…	Good Kenyan girls become
14	Fast Color	100 min	Drama|Sci-Fi|Thriller	Julia Hart	Gugu Mbatha-Raw|Lorraine Toussaint|David Str…	A woman is forced to go on
15	Stuck	90 min	Drama|Musical	Michael Berry	Giancarlo Esposito|Amy Madigan|Ashanti|Arden…	An original pop musical film
16	Shazam!	132 min	Action|Adventure|Comedy|Fantasy	David F. Sandberg	Zachary Levi|Mark Strong|Asher Angel|Jack Dyl…	We all have a superhero in
17	Dumbo	112 min	Adventure|Family|Fantasy	Tim Burton	Colin Farrell|Michael Keaton|Danny DeVito|Eva…	A young elephant; whose c
18	Captain Marvel	123 min	Action|Adventure|Sci-Fi	Anna Boden|Ryan Fleck	Brie Larson|Samuel L. Jackson|Ben Mendelsoh…	Carol Danvers becomes on

3. Movie context information from Twitter

From IMDB, we can get all the new movie titles, use these names as Twitter hashtags names to get the tweet of those movies, get retweet number, time of tweet, etc.

Twitter hashtag

news_id	user_name	screen_name	time_of_tweet	retweet_count	news_tweet	hashtags
1118207194766331905	IMDb	IMDb	Tue Apr 16 17:39:05 +0000 2019	259	❤️ #TuesdayMotivation #HeathLedger @hitRECordJoe #10ThingsIHateAboutYou https:/…	10ThingsIHateAboutYou
1121085188677570561	AMC Theatres	AMCTheatres	Wed Apr 24 16:15:12 +0000 2019	8	Here's the official poster for #21Bridges! Starring @ChadwickBoseman and produced by…	21Bridges
1121124616049553408	Fandango	Fandango	Wed Apr 24 18:51:52 +0000 2019	31	Chadwick Boseman's first movie he's led since #BlackPanther is #21Bridges! a cop thrill…	21Bridges
1121130082771202048	Rotten Tomatoes	RottenTomatoes	Wed Apr 24 19:13:35 +0000 2019	119	The first trailer for @chadwickboseman and @Russo_Brothers new film #21Bridges drop…	21Bridges
1121398454230638592	Fandango	Fandango	Thu Apr 25 13:00:00 +0000 2019	41	As his leading man follow-up to #BlackPanther! Chadwick Boseman closes down the #2…	21Bridges
1121398469581627397	AMC Theatres	AMCTheatres	Thu Apr 25 13:00:04 +0000 2019	18	Here's the first official trailer for #21Bridges starring Chadwick Boseman & produce…	21Bridges
1121398704425062407	Rotten Tomatoes	RottenTomatoes	Thu Apr 25 13:01:00 +0000 2019	344	Chadwick Boseman locks down New York City in the first trailer for the Russo brothers n…	21Bridges
1121398975133835264	IMDb	IMDb	Thu Apr 25 13:02:04 +0000 2019	40	.@chadwickboseman seeks redemption in the new trailer for #21Bridges. https://t.co/aM…	21Bridges
1121400987430920192	Hollywood Reporter	THR	Thu Apr 25 13:10:04 +0000 2019	9	Watch: @chadwickboseman faces killers in the first tralier for #21Bridges https://t.co/X8t…	21Bridges
1121423715693146113	Marcus Theatres 🍿	Marcus_Theatres	Thu Apr 25 14:40:23 +0000 2019	1	The only way out is through him. Experience the new trailer for #21Bridges starring @ch…	21Bridges
1112474711831793666	DreamWorks Anima…	DWAnimation	Sun Mar 31 22:00:14 +0000 2019	23	Varvatos Vex is not pleased that this weekend is so short. @talesofarcadia #3Below htt…	3Below
1113546837175218176	DreamWorks Anima…	DWAnimation	Wed Apr 03 21:00:29 +0000 2019	17	Attention! citizens of Arcadia Oaks! It's National Walking Day! Commence your walk AT…	3Below
1116415661956313088	DreamWorks Anima…	DWAnimation	Thu Apr 11 19:00:10 +0000 2019	30	Who's a good alien puppy?? #3Below #NationalPetDay @talesofarcadia #DreamWorks…	3Below
1118156269762830339	Marcus Theatres 🍿	Marcus_Theatres	Tue Apr 16 14:16:43 +0000 2019	2	This #5DollarTuesday! we have the best deals around! We even have FREE popcorn for…	5DollarTuesday
1077322530921332736	Movies Anywhere	movies_anywhere	Mon Dec 24 21:58:01 +0000 2018	1	Carefull 👀 are fra-gee-lay! Throw on a bunny suit! light up that leg lamp! and sink into th…	AChristmasStory
1120319108044562435	AMC Theatres	AMCTheatres	Mon Apr 22 13:31:04 +0000 2019	7	Happy #EarthDay! Celebrate by seeing #Disneynature's Penguins now playing in @IMA…	AMCTheatres
1120691638823559168	AMC Theatres	AMCTheatres	Tue Apr 23 14:11:22 +0000 2019	101	Two trailers in one day! This is a GREAT #TrailerTuesday. Here's your final trailer for #G…	AMCTheatres
1121202426856427520	AMC Theatres	AMCTheatres	Thu Apr 25 00:01:04 +0000 2019	1	She wants your children. See the terrifying @lalloronamovie now playing at #AMCTheatr…	AMCTheatres
1121216284065247234	AMC Theatres	AMCTheatres	Thu Apr 25 00:56:07 +0000 2019	8	OH SNAP! It's almost time for #AMCWTC. And you KNOW we're headed to the MCU for…	AMCWTC
1121217814063456256	AMC Theatres	AMCTheatres	Thu Apr 25 01:02:12 +0000 2019	33	One more thing! Let's be kind! let's have fun! use #AMCWTC and... https://t.co/Zw9q7AT…	AMCWTC

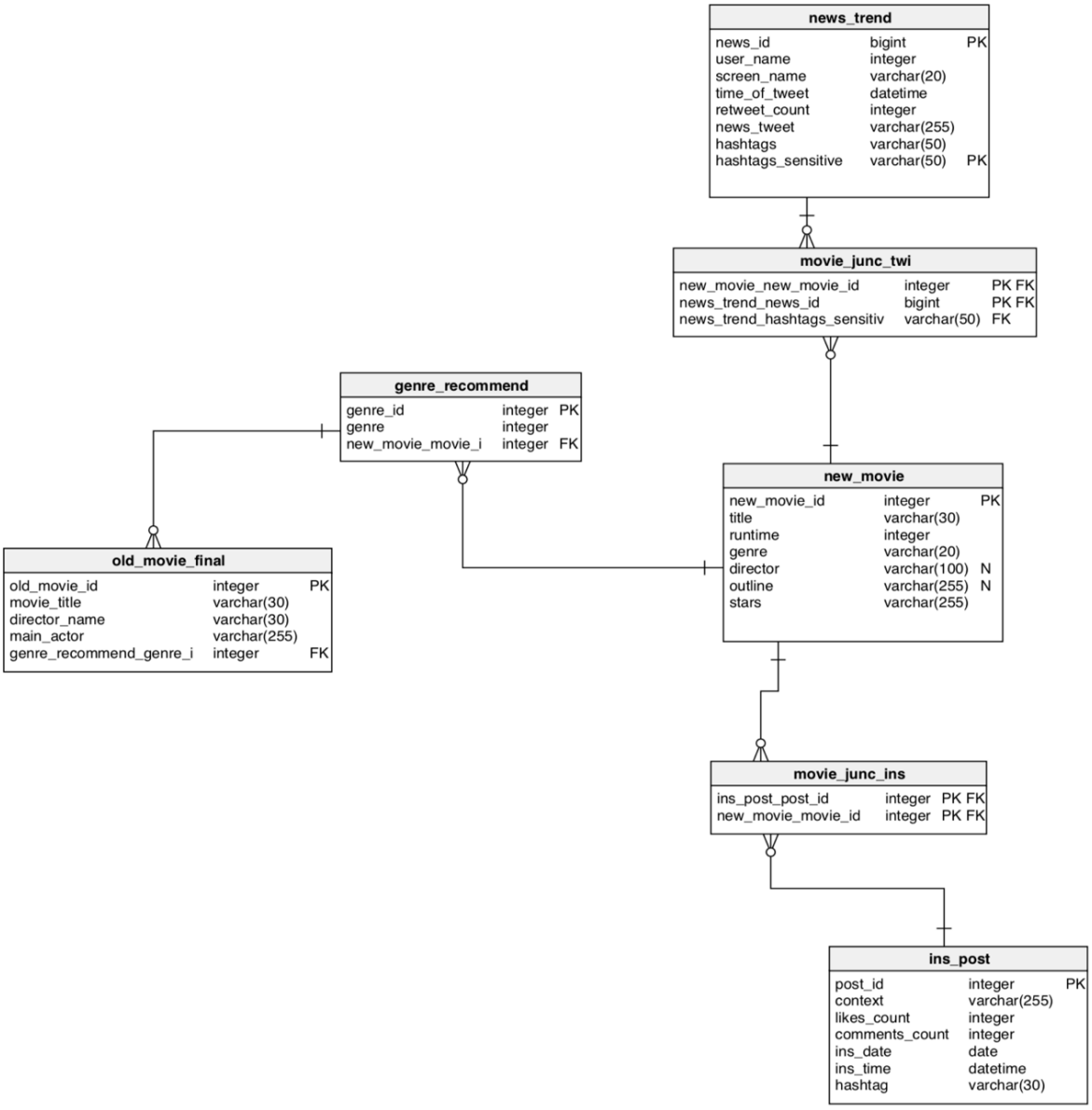
4. Movie fans information from Instagram

We use new movies name got from IMDB as hashtag and we use instaloader to get ins information those post posts in specific hashtags. In order to simplify data, we just extract those who posts and get posts likes, post comments and so on. And there might be some related hashtags.

Inshashtag

post_id	ins_date	ins_time	context	like_count	comments_count	hashtags
1	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	penguins
2	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	disneynature
3	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	documentary
4	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	nature
5	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	animals
6	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	movies
7	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	film
8	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	youtuber
9	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	smallyoutuber
10	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	lgbt
11	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	lgbtq
12	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	gay
13	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	instagay
14	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	gaysofinstagram
15	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	bi
16	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	biboy
17	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	biguys
18	2019-04-17	23:10:46	Seeing DisneyNature's Penguins in IMAX! Always a tradition for me t…	10	0	cute

III. ER Diagram



IV. Code with Documentation

Python browser

Run

```
In [*]: browser()
        Please input a Movie you love :
```

<<Sunshine>> is the movie I love most, we input it into the browser

Run

```
In [*]: browser()
        Please input a Movie you love : Sunshine
```

We get the result <<Captain Marvel>> which has the same genre with <<Sunshine>>.

Run

```
In [59]: browser()
         Please input a Movie you love :Sunshine
         Captain Marvel

         ----------Ins Comments----------
         ('#shazam was a fantastic watch!  Really enjoyed the clever writing that doesn't undermine an audiences knowledge of
         devices and tropes! Did this piece for SDCC and has more relevance today then when it was done a couple years back! G
         O SEE #captainmarvel ERRRRE #shazam',)
         ('\u202aAt long last we have not one but THREE #movie reviews on the #podcast today, as I am joined by @ThyKingdomKam
         to discuss #CaptainMarvel, #UsMovie, and #Shazam. #Marvel #DC and #horror all come together in this episode, link in
         bio or your fav #podcasting platform.\n.\n.\n#podernfamily \u202c#avengers #justiceleague #mcu #dceu #marvelcomics #d
         ccomics #love #video #gif #audio #geek #nerd',)
         ('New vlog is up !\nLink in bio ✨ \n#youtube #youtuber #smallyoutuber #vlog #vlogger #couplevideos #couplevlog #capta
         inmarvel #moviedate #carridechronicles #carride #daywithmyboyfriend #promoter #subscribe #sub4sub #like #comment #sha
         re',)
         ('🌟🌟🌟 \nW A R R I O R.\n#captainmarvel #funkopop #marvel #avengers',)
         ('Part 2\n7. Daredevil #25 [2013] \n8. All-New Ultimates #3 [2014]\n9. Shadowman #8 [2013]\n10. The Punisher War Jour
         nal #80 [1995]\n11. Sub-Mariner #57 [1973]\n12. Captain Marvel #1 [1968]\n\n#marvel #comics #dc #dccomics #valiant #s
         hadowman #issue #read #captainmarvel #60s #70s #marvelnow #ultimates #daredevil #mattmurdock #punisher #warjournal #f
         rankcastle #namor #submariner',)
         ('#shazam was a fantastic watch!  Really enjoyed the clever writing that doesn't undermine an audiences knowledge of
         devices and tropes! Did this piece for SDCC and has more relevance today then when it was done a couple years back! G
         O SEE #captainmarvel ERRRRE #shazam',)
         ('The End Game Sound Track |*Wich is your favorite #hero? Comment bellow and listen a fragment of the soundtrack and
         rise the hype with the animation of Rocket the Racoon. Repost on your profile and support me making mention of my acc
         ount.👍  *¿Cuál es tu héroe favorito? Coméntanos y escucha un fragmento de la banda sonora y aumenta la fiebre con la
         animación de Rocket El Mapache. Repostea en tu perfil y dame los créditos mencionando mi cuenta. 👍  \n#AvengeTheFalle
         n #Avengers #April26 #EndGame #BlackWidow #IronMan #CaptainAmerica #Hulk #AntMan #StarLord #Thor #IronPatriot #Captai
         nMarvel #Gamorra #Thor #CountDown #Style #Happy #FollowxFollow #TagsForLikes #FollowMe #Webstagram #InstaFollow #Phot
         o #InstaLove #Design #Cute #SoundTrack #TheAvengers #RocketTheRacoon',)

         ----------Twitter News----------
         ('Q1: What was the first #Marvel movie you saw at AMC Theatres? #AMCWTC ',)
         ('Q2: Who is your favorite #Marvel character and why? (reply with a gif) #AMCWTC ',)
         ('Q3: If you could bring back one #Marvel character who would it be? #AMCWTC https://t.co/wXi9WpNmM7 ',)
         ("#CaptainMarvel's post-credit scene. Are you ready for #AvengersEndgame?https://t.co/ONm1baMpwM ",)
```

V. News and Trend

The Data for news trend is collected from twitter, I make sure the posts are related to movie by collecting data only from movie news account or movie company account. And I do some following steps to make sure every time I refresh the code, the feedback is the news(posts) about the movie which are in-theater right now:

1. Use web scraper to get all the in-theater movies;
2. Collected all the latest 100 posts of each news account, manage those posts using their hashtags.
3. Using hashtags to match the news with the in-theater movies.

At the time I am writing this example, the in-theater movies are as follows:

```
['Penguins',
 'Breakthrough',
 'HailSatan',
 'LittleWoods',
 'HighontheHog',
 'PetSematary',
 'Us',
 'TheBestofEnemies',
 'Unplanned',
 'HowtoTrainYourDragonTheHiddenWorld',
 'TheCurseofLaLlorona',
 'Kalank',
 'UndertheSilverLake',
 'Rafiki',
 'FastColor',
 'Stuck',
 'Shazam',
 'Dumbo',
 'CaptainMarvel',
 'FiveFeetApart',
 'WonderPark']
```

The news examples are as follow:

1114996656506847234	IMDb	IMDb	Shazam	Sun Apr 07 21:01:33 +0000 2019	32	How well does @ZacharyLevi
1114995758502039552	IMDb	IMDb	IMDbMe	Sun Apr 07 20:57:59 +0000 2019	9	IMDb has been nominated fo
1114966155301666817	IMDb	IMDb	Gladiator	Sun Apr 07 19:00:21 +0000 2019	353	His name is #Gladiator. ⚔ W
1114966155301666817	IMDb	IMDb	RussellCrowe	Sun Apr 07 19:00:21 +0000 2019	353	His name is #Gladiator. ⚔ W
1118627744235921408	20th Century	20thcenturyfox	Gladiator	Wed Apr 17 21:30:11 +0000 2019	7	.@SeeBreakthrough is now pl
1118627744235921408	20th Century	20thcenturyfox	RussellCrowe	Wed Apr 17 21:30:11 +0000 2019	7	.@SeeBreakthrough is now pl
1118554215603793921	20th Century	20thcenturyfox	BreakthroughMovie	Wed Apr 17 16:38:01 +0000 2019	14	Love. Hope. Inspiration. #Bre

Each news(post) is tagged by at least a tag. And when we join these 2 tables, we get all Latest news related the movie.

VI. Use case to answer those question

1.What are people saying about me (somebody)?

```
1    #people's comments about me(black)
2 •  SELECT context FROM ins_post WHERE hashtags LIKE 'black'
```

100% 57:2

Result Grid Filter Rows: Search Export:

context
... #666 #baphomet #black #blackmagic #blasphemy #beast #cult #da…
Omfg😻😻..he is so precious • • • ignore tags:😍 #colesprouse #cole…

2.How viral are my post?
6885 retweet, so viral!!

```
4    #maximum retweet number
5 •  SELECT  max(retweet_count) AS viral_post FROM news_trend
```

100% 57:5

Result Grid Filter Rows: Search Export:

viral_post
6885

3.What posts are likely to be interesting to me?

```
6
7    #The movies have the actor Chris Evans are likely to be interesting to me
8 •  SELECT movie_title FROM old_movie_final WHERE main_actor LIKE 'Chris Evans'
```

100% 75:8

Result Grid | Filter Rows: Search | Export:

movie_title
TMNT
Sunshine
Fantastic 4: Rise of the Silver Surfer

4.What posts are like mine?

```
9
10   #news_tweet has the word 'awesome' like mine
11 •  SELECT * FROM news_trend WHERE news_tweet LIKE '%awesome%'
```

100% 59:11

Result Grid | Filter Rows: Search | Edit: | Export/Import:

news_id	user_name	screen_name	time_of_tweet	retweet_count	news_tweet
1121415897237917697	DR Movie News	DRMovieNews1	Thu Apr 25 14:09:19 +0000 2019	13	Today! the #RoadToEndgame comes to an end...!Join us as
1119269267072061441	Marcus Theatres	Marcus_Theatres	Fri Apr 19 15:59:22 +0000 2019	3	This is incredible! way to go! Steve! I'm sure our #Dreamlour
1117940670969405445	Marcus Theatres	Marcus_Theatres	Tue Apr 16 00:00:00 +0000 2019	1	We've spent the last few days soaking up all things @StarW
1121415897237917697	DR Movie News	DRMovieNews1	Thu Apr 25 14:09:19 +0000 2019	13	Today! the #RoadToEndgame comes to an end...!Join us as
1117940670969405445	Marcus Theatres	Marcus_Theatres	Tue Apr 16 00:00:00 +0000 2019	1	We've spent the last few days soaking up all things @StarW
1099004286334164993	Animation Addict	animationaddict	Fri Feb 22 17:53:35 +0000 2019	9	Finaled my last shot on #ToyStory4 today! thank you @Cool
1099004286334164993	Animation Addict	animationaddict	Fri Feb 22 17:53:35 +0000 2019	9	Finaled my last shot on #ToyStory4 today! thank you @Cool
NULL	NULL	NULL	NULL	NULL	NULL

5.What users post like me?

```
12
13    #posts that time_of_tweet on Saturday are like mine
14 •  SELECT user_name FROM news_trend WHERE time_of_tweet LIKE '%sat%' OR '%wed%'
```

100% 61:14

Result Grid | Filter Rows: Search | Export:

user_name
Rotten Tomatoes
Fandango
Fandango
IMDb
Movies Anywhere
Animation Addict
AMC Theatres
20th Century Fox
AMC Theatres

news_trend 20

6.Who should I be following?

```
16    #I'll follow the account that retweet_count number >500
17 •  SELECT user_name FROM news_trend WHERE retweet_count >500
```

100% 58:17

Result Grid | Filter Rows: Search | Export:

user_name
Rotten Tomatoes
Rotten Tomatoes
DreamWorks Ani…
Warner Bros.
Rotten Tomatoes
IMDb
IMDb
Entertainment W…
Entertainment W…
Rotten Tomatoes

news_trend 21 | Read Only

Result Grid | Form Editor

Action Output

Time	Action	Response	Duration / Fetch Time

7.What topics are trending in my domain?

```
19    # 7 Number of hashtag that mentioned most is trending
20 •  SELECT hashtags as title,COUNT(*) as number FROM ins_post GROUP BY hashtags
21
22
23
```

100% 1:21

Result Grid | Filter Rows: Search | Export: | Fetch rows:

title	number
penguins	5
disneynature	2
documentary	3
nature	3
animals	2
movies	4
film	3
youtuber	2
smallyoutuber	2
lgbt	1
lgbtq	1
gay	1
instagay	1
gaysofinstagram	1
bi	1
biboy	1
biguys	1

```
19    # 7 Number of hashtag that mentioned most is trending
20    #SELECT hashtags AS title,COUNT(*) AS number FROM ins_post GROUP BY hashtags
21    #SELECT MAX(numbers) FROM trendinghashtag
22    SELECT * FROM trendinghashtag WHERE numbers = 6
23
```

100% 1:22

Result Grid Filter Rows: Search Export:

title	numbers
disney	6
petsematary	6
love	6

8.What keywords/hashtags should I add to my post?

```
1    #8
2    SELECT title,hashtag,retweet_count
3    FROM title_hashtags,new_movie,news_trend
4    WHERE title_hashtags.movie_title=new_movie.title
5    AND title_hashtags.hashtag=news_trend.hashtags
```

100% 3:1

Result Grid Filter Rows: Search Export:

title	hashtag	retweet_count
Breakthrough	Breakthrough	14
Breakthrough	BreakthroughMovie	16
Breakthrough	BreakthroughMovie	45
Breakthrough	BreakthroughMovie	7
Breakthrough	BreakthroughMovie	23
Breakthrough	BreakthroughMovie	19
Breakthrough	BreakthroughMovie	11
Breakthrough	BreakthroughMovie	11
Breakthrough	BreakthroughMovie	20
Breakthrough	BreakthroughMovie	11
Breakthrough	BreakthroughMovie	4

9.Should I follow somebody back?

```
25    #9
26    SELECT * FROM user_data WHERE number_of_tweets_in_24!=0 AND followers_num>100
27
```

100% 70:26

Result Grid Filter Rows: Search Export:

user_name	screen_name	followers_num	number_of_tweets_in_24	friends_count
mariana ll-//	miraculousyeet	137	63	822
A loves MBB	brownspoetics	378	19	1616
ㄷH구	httyd_love_	372	18	111
Ryan Wooledge	Ryan26514438	179	13	1700
Nor Cal Mythos	NorCalMythos	4098	52	4438
Cynthia	MixTape_Vol_1	568	106	967
Chiefsheart l AU Hiccstrid	HaruTomoio	935	45	879
Wani	LunziKawaihime	104	4	662

VII. Fuzzy search

Basically two packages are used to build these tag tables, one is difflib, another one is fuzzywuzzy.

Difflib can find similar word with a given string from a given list;

Fuzzywuzzy can value the similar rate of two strings

1. Synonyms

Code:

1. synonyms_hashtag

In [38]:

```
filename = "title_hashtags_all_synonyms.csv"
f = open(filename,"w")

headers = "hashtag,synonyms_hashtag\n"
f.write(headers)

for l in hashtag1:
    h=l.replace(' ','').replace(':','').replace('!','').replace('?','')
    match1=difflib.get_close_matches(h, hashtag1) #Avenger=avenger=avenger4
    matching = [s for s in hashtag1 if h in s] #AvengerEndgame=Avenger
    match2 = list(OrderedDict.fromkeys(matching))
    match_raw=match1+match2
    match = list(OrderedDict.fromkeys(match_raw)) # delete duplicates

    for m in match:
        if m != h:
            f.write(l + "," + m + " \n")

f.close()
```

result (5 examples are shown)

3. synonyms tag

In [39]:

```
df_synonyms = pd.read_csv('title_hashtags_all_synonyms.csv', header=0,low_memory=False)
df_synonyms.head(5)
```

Out[39]:

	hashtag	synonyms_hashtag
0	BigLittleLies	LittleMovie
1	WandaVision	DeWandaWise
2	Avengers	AvengersEndgame
3	Avengers	TheAvengers
4	Avengers	AvengersEndgame

2. Mis-spellings

Basically it's finding word that similar rate >90%, the value 90% is after several trys I found it as a perfect value for check mis-spelling word

Code:

2. spelling mistake or very similar words

In [34]:
```
from fuzzywuzzy import fuzz
from fuzzywuzzy import process
#this package can values differences between strings
```

In [32]:
```
filename = "title_hashtags_all_spell.csv"
f = open(filename,"w")

headers = "hashtag,misspelling_hashtag\n"
f.write(headers)

for l in hashtag1:
    h=l.replace(' ','').replace(':','').replace('!','').replace('?','')
    l2=[]
    match1=difflib.get_close_matches(h, hashtag1) #Avenger=avenger=avenger4
    matching = [s for s in hashtag1 if h in s] #AvengerEndgame=Avenger
    match2 = list(OrderedDict.fromkeys(matching))
    match_raw=match1+match2
    match = list(OrderedDict.fromkeys(match_raw)) # delete duplicates

    for m in match:
        #package fuzz can value the similar rate of 2 string
        #I tried many rate, and found 90% similar rate is apropriate to find spelling mistake
        if fuzz.ratio(m,h)>=90:
            if m != h:
                f.write(l + "," + m + " \n")

f.close()
```

Result (5 examples are shown)

2. mis-spelling tag

In [35]:
```
df_spell = pd.read_csv('title_hashtags_all_spell.csv', header=0,low_memory=False)
df_spell.head(5)
```

Out[35]:

	hashtag	misspelling_hashtag
0	Deadpool2	Deadpool
1	Mavel	Marvel
2	Deadpool	Deadpool2
3	LitteMovie	LittleMovie
4	BreatkthroughMovie	BreakthroughMovie

3. Semantic information

First I have a list of movie titles, a list of TV series, a list of all hashtags. From all hashtags list, those who are 70% similar to movie titles are movies title tags;

From all hashtags list, those who are 70% similar to TV series are TV series title tags;
The rest are in other category

find tags that are similar to movie titles, write it in csv

In [361]:
```
filename = "title_hashtags_all_semantic.csv"
f = open(filename,"w")

headers = "category,semantic_hashtag\n"
f.write(headers)



for l in df_movie_list:
    h=l.replace(' ','').replace(':','').replace('!','').replace('?','')
    match1=difflib.get_close_matches(h, hashtag1)
    # package 'difflib' can get all similar words as a string from a list
    #Avenger=avenger=avenger4
    matching = [s for s in hashtag1 if h in s] #AvengerEndgame=Avenger
    match2 = list(OrderedDict.fromkeys(matching)) # delete duplicates
    match_raw=match1+match2 #add all the possible matches
    match = list(OrderedDict.fromkeys(match_raw)) # delete duplicates

    for m in match:
        f.write('movie_title' + "," + h + " \n")
```

find tags that are similar to TV titles, write it in csv

In [362]:
```
df_tv = pd.read_csv('tv_title.csv', header=0,low_memory=False)
list_tv=df_tv['tv_title'].tolist()
```

In [363]:
```
for l in list_tv:
    h=l.replace(' ','').replace(':','').replace('!','').replace('?','')
    match1=difflib.get_close_matches(h, hashtag1)
    # package 'difflib' can get all similar words as a string from a list
    matching = [s for s in hashtag1 if h in s]
    match2 = list(OrderedDict.fromkeys(matching))# delete duplicates
    match_raw=match1+match2
    match = list(OrderedDict.fromkeys(match_raw)) # delete duplicates

    for m in match: #write all the match result into a csv file
        f.write('TVseries_title' + "," + h + " \n")
```

tags that are not similar to any list I've got

In [364]:
```
list_all=df_movie_list+list_tv
len(list_all)
```

Out[364]: 70

In [365]:
```
list_all_add=[]
```

In [366]:
```
for l in list_all:
    h=l.replace(' ','').replace(':','').replace('!','').replace('?','')
    match1=difflib.get_close_matches(h, hashtag1)
    matching = [s for s in hashtag1 if h in s]
    match2 = list(OrderedDict.fromkeys(matching))
    match_raw=match1+match2
    match = list(OrderedDict.fromkeys(match_raw)) #match is all the similar tag
    for m in match:

        if fuzz.ratio(m,h)>=70:
            list_all_add.append(m)
```

In [367]:
```
len(list_all_add)
```

Out[367]: 36

In [368]:
```
list_all=list_all+list_all_add
```

In [369]:
```
len(list_all)
```

Out[369]: 106

In [370]:
```
list_all
```

Out[370]:
```
['TheWhiteCrow',
 'PetSematary',
 'Us',
 'TheBestofEnemies',
 'Unplanned',
 'HowtoTrainYourDragon',
 'Avengers',
 'BeNatural',
 'Shazam',
 'Dumbo',
 'CaptainMarvel',
 'FiveFeetApart',
 'WonderPark',
 'LongShot',
 'UglyDolls',
 'TheIntruder',
 'ElChicano',
 'MeetingGorbachev',
 'Bolden',
```

In [371]:
```
for l in list_all:
    for h in hashtag1:
        if l!=h:
            f.write('Other_tags' + "," + h + " \n")
```

In [372]:
```
f.close()
```

Result

4. semantic

In [399]:
```
df_semantic = pd.read_csv('title_hashtags_all_semantic.csv', header=0,low_memory=False)
df_semantic.sample(30)
```

Out[399]:

	category	semantic_hashtag
466	Other_tags	STUBER
237	Other_tags	BlizzardLayoffs
24	movie_title	Godzilla
433	Other_tags	HappyNewYear
164	Other_tags	ad
29	TVseries_title	ShortyAwards
431	Other_tags	TheHateUGive
140	Other_tags	GeminiMan
72	Other_tags	TonyAwards
171	Other_tags	Godzilla
214	Other_tags	CloakAndDagger
434	Other_tags	ToyStory
9	movie_title	LongShot

4. tags in my domain

All the tag in domain movie

Code

4. tags in my domain

In [391]:
```
filename = "title_hashtags_all_domain.csv"
f = open(filename,"w")

headers = "domain,hashtags\n"
f.write(headers)

for l in df_movie_list:
    h=l.replace(' ','').replace(':','').replace('!','').replace('?','')
    match1=difflib.get_close_matches(h, hashtag1) #Avenger=avenger=avenger4
    matching = [s for s in hashtag1 if h in s] #AvengerEndgame=Avenger
    match2 = list(OrderedDict.fromkeys(matching))
    match_raw=match1+match2
    match = list(OrderedDict.fromkeys(match_raw)) # delete duplicates

    for m in match:
        f.write('movie' + "," + h + " \n")
f.close()
```

Result

1.Domain tags

In [397]:
```
df_domain3 = pd.read_csv('title_hashtags_all_domain.csv', header=0,low_memory=False)
df_domain3.head(5)
```

Out[397]:

	domain	hashtags
0	movie	TheWhiteCrow
1	movie	Us
2	movie	TheBestofEnemies
3	movie	HowtoTrainYourDragon
4	movie	Avengers

VIII. Result90% by ourseves

We create the database about latest movies and related old movies information. Then we write code in python to clean data, and then put into SQL to build connection of our database. We write functions to help users to find the movie they would like to watch.

IX. Conclusion and Difficulties

We create the database and use python to clean and then build connection of our tables and information. Then we write switch case to call database. Therefore, when user search information, we will call function to get the information they want.

However, the most difficult thing is that we don't know how to create a website. And there are limits of using twitter API. So, from this project, we have clear aim of what to learn and do in the future: consolidate our knowledge, try to learn how to create the website.

X. Citation

90% by ourselves

10% by external sources

1. Kaggle: https://www.kaggle.com/beaubellamy/ski-resort

2. Instagram API: https://instaloader.github.io/index.html
3.Twitter API: https://developer.twitter.com/en/docs.html
4.http://www.omdbapi.com/
5.https://www.themoviedb.org/documentation/api
6.https://github.com/NIKBEARBROWN/INFO_6210

MIT License

LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.