

Flamingo Forensics Data Science Technical Challenge

This document includes two parts:

1. The answers for Task 1 part A-D
2. Explanation on the methodology of completing Task 2

Task1

a) What does the “doc” variable store and contain? (2 marks)

Ans: As used to store the return from the Jsoup *get()* command, “doc” stores all the information in HTML format from the website.

b) What does the “el” variable contain and why is this line of code useful for similar extraction tasks? (4 marks)

Ans: by using *getElementById()*, the *el* variable searches and stores the element from “doc” that contains the ID attribute “QuoteTextID” .

This code is helpful in a sense that it can target on the specific element inside a website, since the ID used in a page is unique.

c) What does the “els” variable contain? Why is this variable missing other relevant extraction results? (4 marks)

Ans: by using *getElementsByClass()*, the *els* variable searches and stores the collection of all elements from “doc” that is under the class name “feed-image” . And The reason why “style” is missing in the extraction is that Jsoup cannot interperate CSS format. To solve the problem, we need another CSS parser to help extract the information.

d) What is the final returnable output of activisionFinder()? Describe how you came to this decision. (5 marks)

Ans: 2. The if condition is true since *el* in string format is not NULL, therefore *els.size() + 1* is returned to the main function. And since *els* elements has only one “div” section, *els.size() + 1* will return 2.

Task2: Identify Top 100 Influencer in Gaming

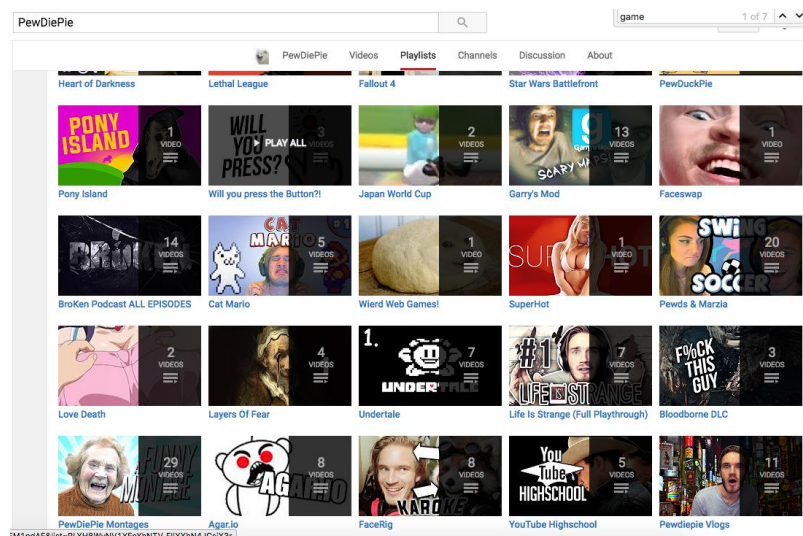
1. PRELIMINARY RESEARCH

1.1 The profiles of the 3 players listed in the technical challenge instructions

To get a sense of what does influencers mean, I did a preliminary research on YouTube based on the examples provided in the challenge instruction pamphlet.

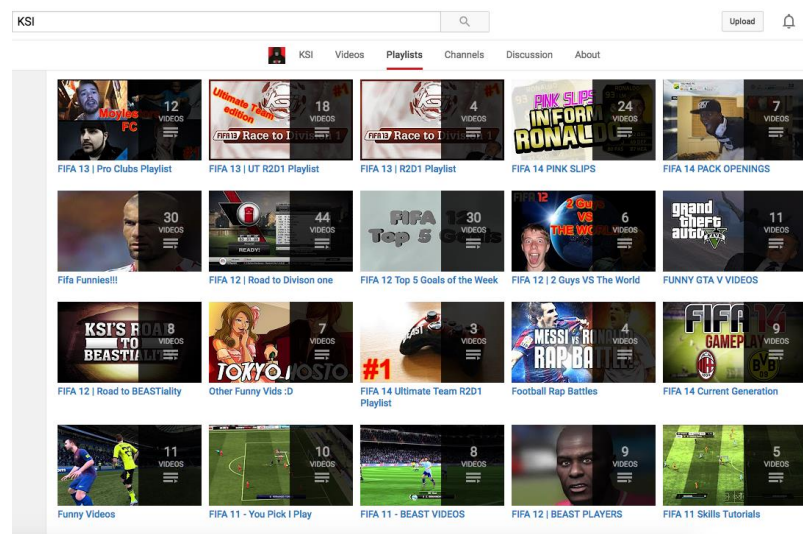
PewDiePie:

Video commentary & Vlogs; average hit: 4,323,560 views



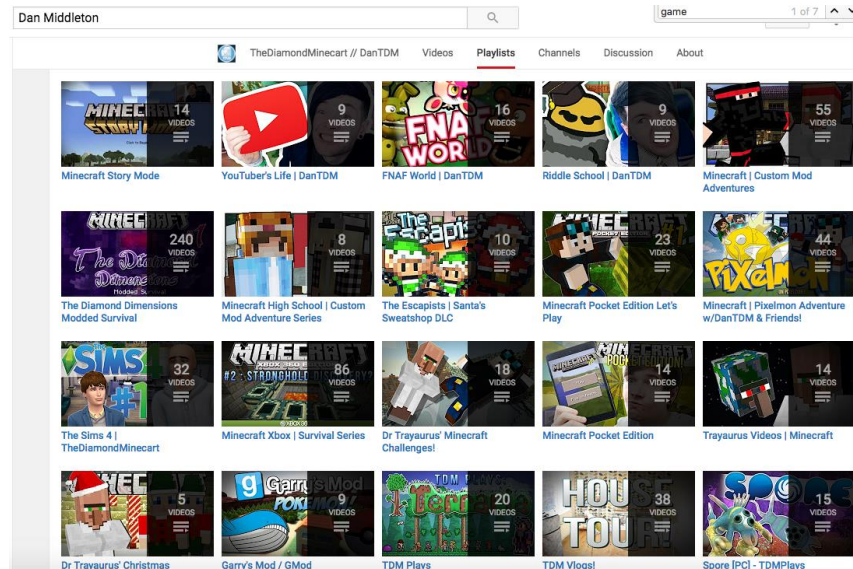
KSI:

FIFA, same time rapper, actor, controversy average hit: 4,354,720 views



Dan Middleton:

Mostly Minecraft, average hit: 1,930,995 views or somewhat



As reflected by the rough survey, influencers stand out and their popularities can be indicated by a number of factors like “the number of views, likes, favorited, shares, and comments etc.” . In particular, it is noticeable that though some people gain fame because of their generalized commentaries or mockery on games, the majority **specialized in a particular type of game**.

Secondly, as observed in YouTube, while other may put their interest or expertise in their description, influencers like PewDiePie does not have any tags or descriptions that could let newcomer to gaming identify him an influencer in **GAMING**.

Because of the two mentioned characteristics, **it is difficult to directly filter the influencer out based on the search of keywords like “Game “or” Gamer “as the direct indication for such channels maintained by the influencers**.

Therefore, I tried to approach in a more indirect way: **from game to gamer**.

1.2 Implementation on the Approach

As the task is focusing on the gaming influencers particularly on social media, my preliminary research tried to source information that would provide me a basic yet holistic landscape of the gaming industry. Making influence/impact as the indicators, I am particularly interested in the following 3 elements:

- Social media platforms
- Game platforms
- Game played on different platforms

1.2.1 Social Media:

From the survey screenshot listed below, it is easy to conclude 3 most heavily used platforms are Facebook, YouTube, and Twitter. They are the three platforms I decided to excavate for this challenge (though I decided to take into account only YouTube and Twitter at the end since Facebook has restrictions on the public access to its Data APIs from a third-party platform).

It is nevertheless remarkable to point out that other platforms like LinkedIn, Pinterest, Instagram, Snapchat or Tumblr are all popular, albeit they are more or less specialized social networking platforms that would serve a particular group of people for particular use, whereas Facebook and Twitter are more ubiquitous and generalized-purposed.

In addition, YouTube is the most popular **video platform** as it is a perfect media for gaming contents to be publicized. However, nowadays more and more online gaming platforms are coming out, like **Stream**, and they are more relevant in a sense that could provide a concentrated dataset for analysis, but because of the time limit and the fact I failed to find the APIs for such websites, I did not include them into the analysis.

Ref: <http://www.ebizmba.com/articles/social-networking-websites>

Top 15 Most Popular Social Networking Sites | August 2016

Here are the top 15 Most Popular Social Networking Sites as derived from our *eBizMBA Rank* which is a continually updated average of each website's *Alexa Global Traffic Rank*, and U.S. Traffic Rank from both *Compete* and *Quantcast*. "*" Denotes an estimate for sites with limited data.



1 | Facebook

3 - eBizMBA Rank | 1,100,000,000 - Estimated Unique Monthly Visitors | 3 - Compete Rank | 3 - Quantcast Rank | 2 - Alexa Rank | Last Updated August 25, 2016.
The Most Popular Social Networking Sites | eBizMBA



2 | YouTube

3 - eBizMBA Rank | 1,000,000,000 - Estimated Unique Monthly Visitors | 4 - Compete Rank | 2 - Quantcast Rank | 3 - Alexa Rank | Last Updated: August 25, 2016.
The Most Popular Social Networking Sites | eBizMBA



3 | Twitter

12 - eBizMBA Rank | 310,000,000 - Estimated Unique Monthly Visitors | 21 - Compete Rank | 8 - Quantcast Rank | 8 - Alexa Rank | Last Updated August 25, 2016.
The Most Popular Social Networking Sites | eBizMBA

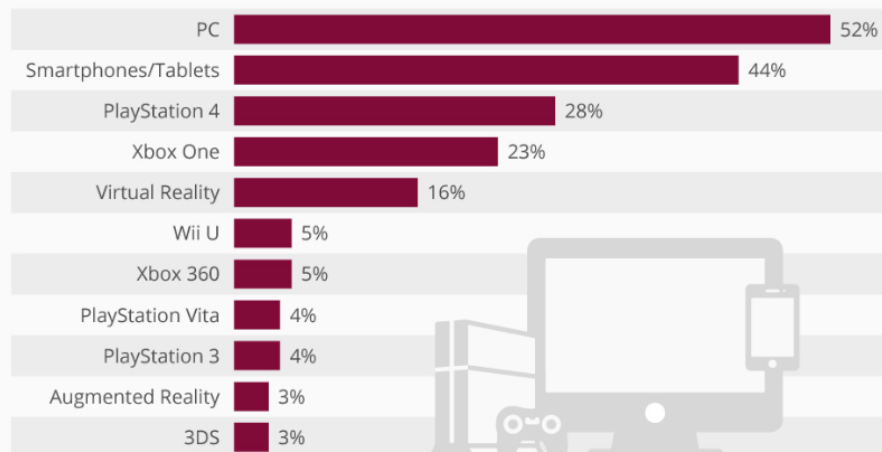
1.2.2 Game Platforms

According to the statistic platform *Statista*, the most popular devices nowadays are: **PC**, **Video Game**: PS, Xbox, **VR**, **Wii** (Nintendo), **Mobile Platform**: iOS, Android

Ref: <https://www.statista.com/chart/4527/game-developers-platform-preferences/>

The Most Important Gaming Platforms in 2016

% of developers who are working on a game for the following platforms



Based on a survey of 2,000 game developers
Source: Game Developers Conference

statista

However, for this challenge, I intentionally left out VR as VR technology is relatively immature providing with its low market penetration rate and high prices. In addition, though mobile games are very popular, itself does not generate much revenue in per-user basis. Moreover, in general people treat mobile games much less serious and they are mainly for relaxation or time-killing purposes, hence providing with such observation, I decided to take only the top 3 mobile game into consideration.

Therefore, in the analysis I will only focus on games played on PC, PS, Xbox, Wii and as mentioned before, the online platform (in this case Stream).

1.2.3 Top Games Played on Different Platforms

A list of game, hence KEYWORD used for searching for this study:

PC: Minecraft, Overwatch, WOW (World of War), CS (Counter-Strike), Diablo, StarCraft

PS: Grand Theft Auto, Gran Turismo, The Last of Us, Uncharted, Final Fantasy

Xbox: Call of Duty, Grand Theft Auto V, FIFA, 2K16, Battlefield

Wii: Nintendo Land, Hyrule Warriors, Pokken Tournament

Stream: Dota 2, Team Fortress 2, Unturned

Mobile: Pokemon Go, Clash of Clans, Candy Crush

Besides that, generalized search word "Gaming" are also added into the search list

Ref: https://en.wikipedia.org/wiki/List_of_best-selling_video_games#PC

https://en.wikipedia.org/wiki/List_of_best-selling_PlayStation_3_video_games

<https://store.xbox.com/en-us/Xbox-One?SortBy=MostPopular>

https://en.wikipedia.org/wiki/List_of_best-selling_video_games#Wii

<http://store.steampowered.com/stats/>

The Top 5 from each platform are selected, but because of the cross-platform issues, lists in Wii and Steam only shows 3 entries.

2. QUANTITATIVE SEARCH DESIGN

2.1 Search Method

By bring in Proxies, I transformed the direct measurement to the influencer to the indirect approach. That is, to firstly search the keyword of the popular games (26 in total) across the 3 social platforms (Facebook, YouTube, Twitter). That brings flaws of course, since the single keyword cannot include all the entries of information we would be interested in.

Specifically, apart from the game and platform issues mentioned before, the factors concerning the accuracy and comprehensiveness of the search are:

- **Influencers' name/ID:** may have different names/ids across different platforms
- **Nickname of a Game/Derivatives of a Game:** Some game has multiple names, or different versions. E.g. StarCraft has 2 versions, and World of War can be shortened to be WOW. This brings difficulties for me as an amateur to the gaming industry, where I cannot exhaust the other alternative names for a particular game, or sometimes impossible, as a search for WOW could be surely meaningless as would include plenty irrelevant "feeling" content, and filtering them out can be potentially time-consuming and hard to implement (as will need to **tokenize the tweet/post** and analyze the content)
- **Byproduct of the game:** Search by keyword do not necessarily include all important information on a particular game (game events for example, *ESL one manila*, is the tournament for data 2, and that may involve numbers of influencer in Dota to participate, but it is highly possible that some tweet about that does not include any keywords even in the slight resemblance of Dota)
- **Records from other Game Communities:** (As mentioned before) some game communities like Stream or Raptr cannot be included in the analysis, even though their impact might be significantly not negligible. They might have another composite of influencer also particularly in gaming industry.
- **Timing Issue:** what is the time frame that is proper for the query (short-term, long-term, real-time streaming)?

These are the potential confounders that could hinder the accuracy of the sampling.

2.2 Performance Metrics and Program Flow

The performance metrics for a single video/tweet/post is quite mixed. For example, to measure the impact of a tweet, the number of retweets or being favorited both could be the indicator of the popularity.

As I am doing search on both YouTube and Twitter, I will be discussing the choice for the performance metrics separately. And integrated in the description is the program flow for my algorithm.

2.2.1 YouTube

The case for YouTube is relatively easy. YouTube Data API for developer is very user-friendly. It not only has a repository of API accessing library for Python, but also includes code examples for different platform on its websites. Moreover, a user in YouTube will possess his own individual channel, and that would make the search easier since the channel will contain all the information for a particular user.

The following words shows for the flow of the program:

Firstly, to retrieve the raw data in JSON format using the google python library as: (sample result: use the example code for Search by keyword, then search Overwatch for only Channel type)

```
{
  "kind": "youtube#searchListResponse",
  "etag": "\"I_8xdZu766_FSaexEaDXtIfEWc0/JYfYxJ3P-CBCZq59kMMmmVPFq_I\"",
  "nextPageToken": "CAUQAA",
  "regionCode": "SG",
  "pageInfo": {
    "totalResults": 1000000,
    "resultsPerPage": 5
  },
  "items": [
    {
      "kind": "youtube#searchResult",
      "etag": "\"I_8xdZu766_FSaexEaDXtIfEWc0/GCT3MMrfuJhhiDSwtliiK8fRDQw\"",
      "id": {
        "kind": "youtube#channel",
        "channelId": "UCIOf1XXinvZsy4wKPAkro2A"
      },
      "snippet": {
        "publishedAt": "2014-09-27T06:22:50.000Z",
        "channelId": "UCIOf1XXinvZsy4wKPAkro2A",
        "title": "PlayOverwatch",
        "description": "Overwatch™ is a highly stylized team-based shooter set in a future worth fighting for. Every match is an intense multiplayer showdown pitting a diverse cast of ...",
        "thumbnails": {
          "default": {
```



```

"url": "https://yt3.ggpht.com/-M2-DH_DN-
Qs/AAAAAAAAAAI/AAAAAAAAAA/kczw_wtvMOI/s88-c-k-no-mo-rj-c0x0xfffff/photo.jpg"
},
"medium": {
"url": "https://yt3.ggpht.com/-M2-DH_DN-
Qs/AAAAAAAAAAI/AAAAAAAAAA/kczw_wtvMOI/s240-c-k-no-mo-rj-c0x0xfffff/photo.jpg"
},
"high": {
"url": "https://yt3.ggpht.com/-M2-DH_DN-
Qs/AAAAAAAAAAI/AAAAAAAAAA/kczw_wtvMOI/s240-c-k-no-mo-rj-c0x0xfffff/photo.jpg"
}
},
"channelTitle": "PlayOverwatch",
"liveBroadcastContent": "upcoming"
}
}
Ref: https://developers.google.com/youtube/v3/code\_samples/python#search\_by\_keyword

```

In the code, I am only extracting the information from the channel_id, page_token, since if parameter is correctly set, the search automatically returns the result sorted by view_count of that particular channel, hence no need to ask for large amount of data and do sort myself. I continued to use the channel_id extracted from all the 26 times of search, combining them and send to the second round of API query asking for channel_subscriber for the channels accordingly. The third round of API query is to extract the channel name (it can be down with the second round also, but in this way memories will be less used). Finally, I sort the channel based on the number of subscriber and return the result. The result will be shown in **part 3**.

To put in short, the performance metric I set for YouTube is simply the **number of subscriber for each channel**.

Please refer to the file [YouTube_final.py](#) in the package for detailed coding.

2.2.2 Twitter

Twitter case is much more complicated in both the conceptual and realization level.

Though on the web there are some open-source library dealing with Twitter API access (I used **tweepy**), Twitter official developer site does not include any example codes, so I did all the coding myself. In addition, Twitter **search_tweet** (information for Tweet and User) and **search_user** (information for user mentioned) method will return a full list of the tweets information, out of which I choose the following 15 sets of data to be recorded:

Tweet	User	User_mentioned
Tweet_ID (tracking the tweet) Tweet_favorite (the favorite count for the tweet) Tweet_retweet (the retweet count for the tweet)	User_ID User_screen_name User_favorite User_follower User_friends User_statuses	•User_mentioned_ID •User mentioned_screen_name •User mentioned_favorite •User mentioned_follower •User mentioned_friends •User mentioned_statuses

In the implementation though, I did not include all the information, but from my initial consideration, they should be grouped in the following way:

- IDs: for tracking purpose
- Screen_names: for display
- Tweet_favorite/Tweet_retweet: as an indicator for the popularity of a tweet
- User/User_mentioned Follower: as a measure for the popularity of the user involved in that tweet
- User/User mentioned Friends/Statuses/Favorite: Indicator of the activeness of user

2.2.2.1 Script Description

The twitter package is made up by two python scripts, *Twitter_one_keyword.py* and *Twitter_two_postprocess.py*. The former one is used to crawl the above-mentioned information down from the twitter API, and the later one is used to do the necessary processing, visualization and analysis.

In detail, *Twitter_one_keyword.py* will get connections with the twitter API and download the information for each of the 26 popular games as keyword. It will from twitter extract approximately 1,000 entries of relevant tweets and write them into separate csv files with name "data_twitter_(game name).csv" as could refer in the package. Since twitter imposes a rate limit for individual to access to its data API, the whole program will finish running in about 6 hours.

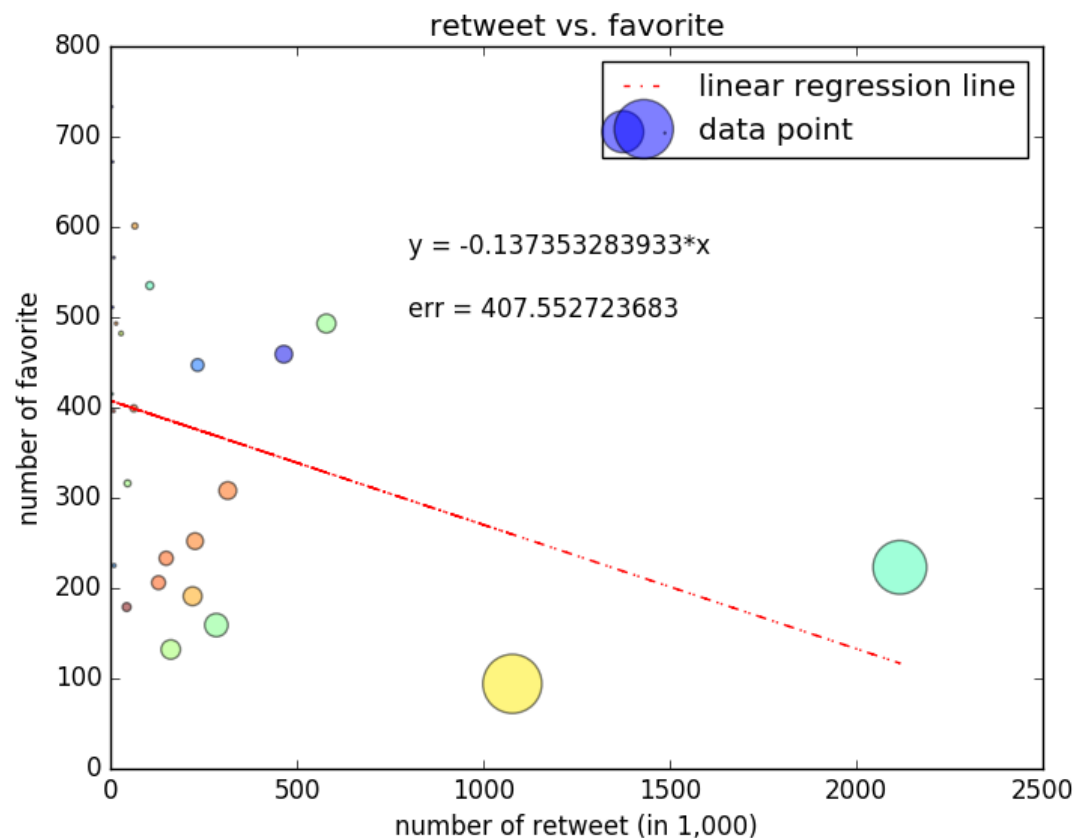
Twitter_two_postprocess.py will then load the data from the 26 files, and there are two metrics set for the measurement which individually takes in different parameters. After processing/sorting, it writes the relevant files to csv format for storage. The last section is more of the appendix as to visualize the relationship between favorite and retweet for a tweet, as to bring testimony why favorite count is not selected as parameter in the processing.

2.2.2.2 Choice of Performance Metrics for the twitter

As described above, two metrics are set separately for the analysis.

Among the above-mentioned 15 pieces of dataset, "friends/statuses/favorite" as the indicator for the activeness of user is dropped for the sake of simplicity. It could be considered at the later stage integrating them to the final output, but it could also bias our judgement as some twitter account may be historically very active.

In addition, the number of retweets is parameterized as the only measurement for the popularity of tweet. Aggregating across the 26 topics, the following scattered plot shows the relationship being a tweet being favorited and retweeted.



The graph clearly shows that the scatter points fail to converge to the linear regression, as the slope is not only close to zero but also negative (therefore counter-intuitive). When examining the data in detail, most tweet, particularly those with low retweet count, has zero favorite count, and that may attribute to the user habit of twitter as they retweet the trending information and favorite the tweet they will look back later. And that intrinsic inconsistency of using these two functions can make the effort to unify the two parameters in vain. Specifically for this research, retweet correlates more with popularity trending, and hence favorite is ignored.

Back to Metrics:

Metrics 1 & 2 are measuring the popularity of individual user in terms of their followers. The user pool consists of both the user who tweets that tweet, and the user mentioned in that tweet. This metrics is relatively straightforward as it uses the absolute follower count for determination.

I named them as "**most_popular_user**" and "**most_mentioned_user**" in the final output.

Metrics 3 & 4 are grouping user and user mentioned in each 1,000 tweets. The program then **sums up** the number of retweets for a particular user as a measure for his influence in that topic; and **count** the number of times the user_mentioned being mentioned as a measure for that person's authority in that topic.

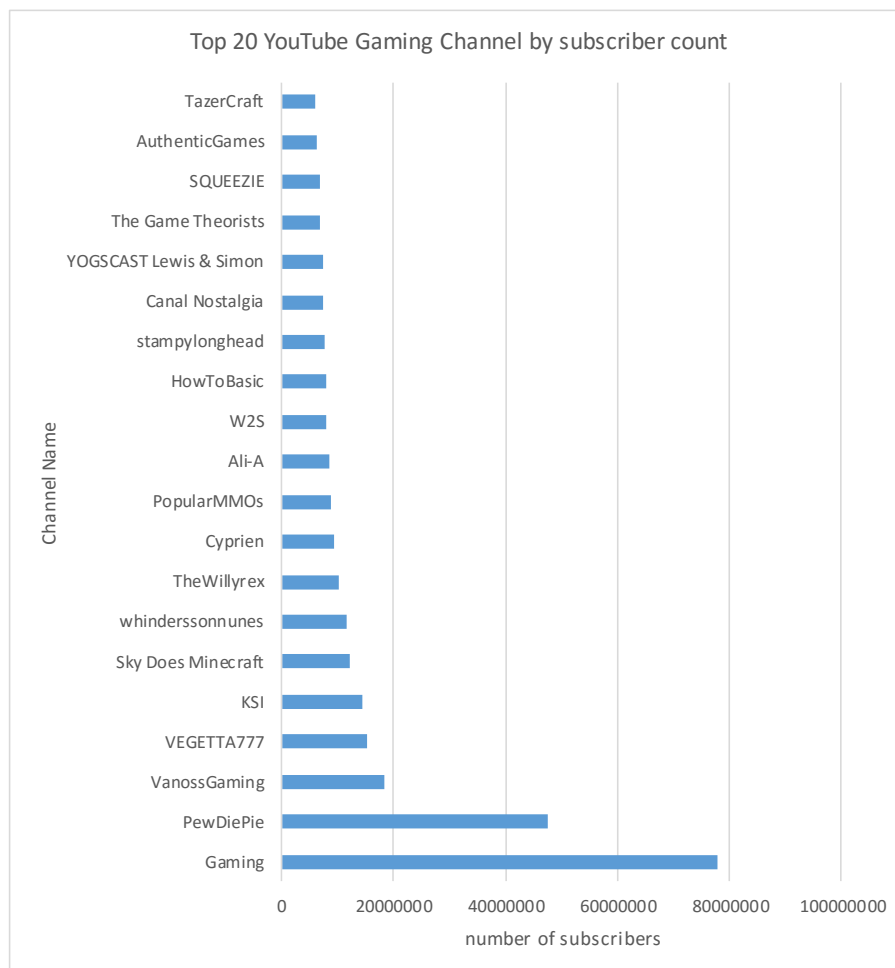
I named them as “**most_recognized_user**” and “**most_recognized_user_mentioned**” in the final output.

The detailed observation would be presented in part 3.

3. Results and Discussions

3.1 YouTube

The graph shows the bar chart for the top 20 most subscribed channel in YouTube from this search.



The complete result dataset is in the csv file [output_youtube.csv](#), it gives the top 100 gaming channel that has the largest number of subscribers.


From the visualization, some observations can be made:

- The YouTube official gaming channel is the most heavily subscribed channel among the survey. This is nevertheless reasonable; I did not exclude it from the result since social media accounts are not necessarily maintained/owned by a single individual. And in reality, official accounts are always popular since there are, authority themselves and publishing their own firsthand stories and updates.

- As a random check from the list, some accounts are not exclusively popular because of its gaming content. Example would be HowToBasic, it was found because of one episode devoted for “How to Catch Rare Pokémon on Pokémon GO” , but this channel has very generalized content and not particularly devoted for gaming. This suggested that for further exploration, we should add a filter that would measure the percent of video contents in a channel that is relating to game. We would reject those whose percentage are below a certain threshold, but how to choose such threshold would need another round of survey to observe some already known game channel and their content in order to average and set such standard.

As a cross-validation, I found an interesting website that publicize the real time gaming ranking in YouTube channels.

Ref: <http://vidstatsx.com/youtube-top-100-most-subscribed-games-gaming-channels>


YouTube Stats (Subscriber, Ranking, & Video Statistics)
Search & Submit

Most Subscribed
Most Viewed
Top Gainers
Top Losses
Future Rank™

YouTube Top 100 Most Subscribed Games & Gaming Channels List - Top by Subscribers

One Hundred Most Subscribed Games & Gaming Channel Rankings List by Subscribers

Video Producer	Report Rank	Subscribers	Sub Rank	24 Hour Sub +/-	7 Day Sub +/-	Videos	Views
PewDiePie	1	47,526,049	4	26,721	166,406	2.9K	13.18 B
VanossGaming	2	18,418,398	18	11,489	70,842	493	5.64 B
VEGETTA777	3	15,096,223	23	11,459	63,222	3.2K	5.37 B
KSI	4	14,346,306	27	27,343	90,729	1K	2.93 B
Markiplier	5	14,220,585	29	9,605	65,572	3.2K	5.47 B
TheDiamondMinecart // DanTDM	6	12,251,313	40	11,141	65,134	2.1K	7.99 B
Sky Does Minecraft	7	12,082,218	42	720	2,549	1.6K	3.35 B
TheSyndicateProject	8	9,958,440	69	592	4,725	3.2K	1.93 B
CaptainSparklez	9	9,519,994	76	7,365	28,247	3.3K	2.59 B
Ali-A	10	8,456,986	100	2,396	14,007	2.2K	2.07 B
rezendeevil	11	8,160,374	110	10,404	50,592	3.6K	2.79 B
W2S	12	7,941,702	113	4,206	32,487	563	1.95 B
H2ODelirious	13	7,774,100	120	8,685	41,654	971	1.22 B
IGN	14	7,608,369	124	2,946	23,444	117.8K	5.64 B
League of Legends	15	7,583,332	125	919	5,027	626	1.24 B
Smosh Games	16	6,936,167	144	2,089	13,803	1.9K	2.16 B
TobyGames	17	6,891,201	147	-305	-1,584	4.8K	2 B
The Game Theorists	18	6,877,734	149	4,031	17,468	218	889.9 M
SQUEEZIE	19	6,722,127	153	10,577	59,200	931	2.61 B
speedyw03	20	6,672,605	157	-127	-299	1.8K	1.48 B
theRadBrad	21	6,298,341	170	2,872	19,128	4.6K	2.29 B
AuthenticGames	22	6,290,730	171	11,661	70,110	2.4K	2.52 B

It seems that my own algorithm caught around 50% in the top 20 count, providing that if this list is accurate.

Moreover, it also suggests other parameters that could be further looked into like the trending factors, and the view count for channels.

3.2 Twitter

Statistics from Twitter is less unified because of results differs when the metrics varies

The following chart shows the Top 20 twitter account from each metrics.

most_popular_user_tweet:			
	user_screen_name	user_follower	Game_name
0	FIFAcorn	9130405	FIFA
1	kompascom	5958495	Gran Turismo
2	EW	5598031	Uncharted
3	bepe20s	5283219	FIFA
4	NintendoAmerica	4946601	Gaming
5	CNNnews18	2720891	The Last of Us
6	diarioas	2166287	FIFA
7	Independent	2009171	The Last of Us
8	MailOnline	1775412	Candy Crush
9	gameinformer	1720432	World of warcraft
10	Variety	1489346	Grand Theft Auto
11	verge	1430866	Call of Duty
12	livemint	931261	Counter Strike
13	Salon	842682	The Last of Us
14	G2A_com	648877	Counter Strike
15	VentureBeat	634833	World of warcraft
16	IIJERiiCH0II	619929	Gaming
17	MotherJones	619860	World of warcraft
18	ATVIAssist	597463	Call of Duty
19	dani3palaciosdj	589092	Diablo

This metrics returned a diverse profile of users. Though through random check the majority are gaming account, some are not, like the easy-to-observe “CNNNew18” ; and “Variety” , which according to its description, an official account from Variety website, which deals with the business of entertainment – films, digitals, TVs in general – and the fact it was involved with WOW topic is that WOW is also the name for the recently released movie.

most_popular_user_mentioned:			
	user_mention_screen_name	user_mention_follower	Game_name
0	rihanna	65008087.0	Battlefield
1	YouTube	63811880.0	Dota2
2	YouTube	63784308.0	Clash of Clans
3	YouTube	63783988.0	Counter Strike
4	YouTube	63782826.0	Candy Crush
5	YouTube	63782171.0	Pokemon Go
6	YouTube	63781947.0	Unturned
7	YouTube	63781646.0	Team Fortress
8	YouTube	63781265.0	Pokken Tournament
9	YouTube	63781112.0	Hyrule Warriors
10	YouTube	63781085.0	Nintendo Land
11	YouTube	63780693.0	Battlefield
12	YouTube	63780556.0	Destiny
13	YouTube	63780467.0	2K16
14	YouTube	63780294.0	FIFA
15	YouTube	63780141.0	Call of Duty
16	YouTube	63780020.0	Final Fantasy
17	YouTube	63779833.0	Uncharted
18	YouTube	63779642.0	The Last of Us
19	YouTube	63779519.0	Gran Turismo

The most obvious observation as YouTube is mentioned across different games. This can also be cross-link to the observation we made for the YouTube search, where the Gaming channel is the most popular subscription in YouTube. Rihanna, on the other hand, is a registered player in Battlefield and at the same time very famous, but obviously not because of her playing the game.

20	YouTube	63779417.0	Grand Theft Auto
21	YouTube	63779291.0	StarCraft
22	YouTube	63779150.0	Diablo
23	YouTube	63778285.0	World of Warcraft
24	YouTube	63778103.0	Overwatch
25	YouTube	63777868.0	Minecraft
26	YouTube	63777610.0	Gaming
27	Harry_Styles	29328953.0	The Last of Us
28	iamsrk	20975715.0	Pokemon Go
29	coldplay	17376852.0	The Last of Us
30	PlayStation	11934046.0	Overwatch
31	WSJ	11680359.0	Grand Theft Auto
32	Xbox	10078464.0	Battlefield
33	lindsaylohan	9308240.0	Grand Theft Auto
34	FIFACOM	9130391.0	FIFA
35	mashable	7682765.0	Pokemon Go
36	mashable	7681719.0	Grand Theft Auto
37	washingtonpost	7184070.0	FIFA
38	EPN	5661952.0	The Last of Us
39	EW	5598023.0	Uncharted

Or we could look at the ranking at the later part, then other participants turned out. Nevertheless, it is obvious that some of them are known to the world but not attribute to game.

most_recognized_user_tweet:			
	user_screen_name	tweet_retweet	Game_name
0	StylesIDforever	149737	The Last of Us
1	larentskid	92504	FIFA
2	Iushgod	92504	FIFA
3	cheluFR	63698	FIFA
4	Criistiina95	63698	FIFA
5	FerJFlores21	63697	FIFA
6	NiFaveo	63697	FIFA
7	emma_cangri	63697	FIFA
8	apliferia	63697	FIFA
9	GwerriorHd	63696	FIFA
10	NeonExeoPvp	63440	Minecraft
11	plowno	63440	Minecraft
12	RevoluitionGam3	63440	Minecraft
13	walkingonadream	46322	The Last of Us
14	LouisMyAttack	46322	The Last of Us
15	Inthearmsofmike	35268	Pokemon Go
16	nutjung_destiny	32437	Destiny
17	cluclu_land	21616	Nintendo Land
18	burntotears	20411	Overwatch
19	LordBeerus23	20411	Overwatch

This metrics again returned a list of diversified user, and it shows that this group of users are the true representative of the gamer. However, there are also flaws with such search method. For example, the user “GwerriorHd” has only one tweet (about FIFA) on his timeline, only has 1 follower, but received over 67K retweets for that tweet. Even if it is not the bug of Twitter, it is hard to categorize such users. Nevertheless in general, this is potentially the best candidate if only single metric be used.

most_recognized_user_mentioned:			
	user_mention_screen_name	user_mention_counts	Game_name
0	gourmetspud	614	Unturned
1	YouTube	448	Minecraft
2	YouTube	343	Team Fortress
3	YouTube	302	Pokken Tournament
4	YouTube	247	Call of Duty
5	YouTube	243	Clash of Clans
6	YouTube	236	Battlefield
7	maximilian_	222	Gaming
8	YouTube	203	Counter Strike
9	ark_akiba	175	Counter Strike
10	YouTube	135	Unturned
11	YouTube	127	2K16
12	YouTube	109	Grand Theft Auto
13	RECfilming	107	Battlefield
14	YouTube	100	World of warcraft
15	Battlefield	98	Battlefield
16	Cdiscount	95	Uncharted
17	YouTube	92	Pokemon Go
18	YouTube	89	Gran Turismo
19	StarCraft	88	StarCraft

The last metrics again shows similar pattern as the second metrics.

Theoretically, as I argued before, 4 metrics all have their own focus and merits and therefore it is difficult to value one more superior than the others. The first two metrics focus on the absolute popularity measured by follower, and does not take into account the relevance of the user to that topic. That is to say, in some cases, a famous user only tweet/being mentioned in only one tweet about game, but would be nevertheless recorded in the list.

The last two metrics is more comprehensive. Measuring for a given time period, the sum of tweet concerning gaming topic being retweeted from one account (Metrics 3), or the count of times being mentioned in a gaming topic (Metric 4) certainly integrate both the popularity and the activeness of the user into evaluation. They differ from each other in terms of their role: Metric 3 gives the list of participant, whereas Metric 4 gives the list of authority, or watcher/observer of that topic.

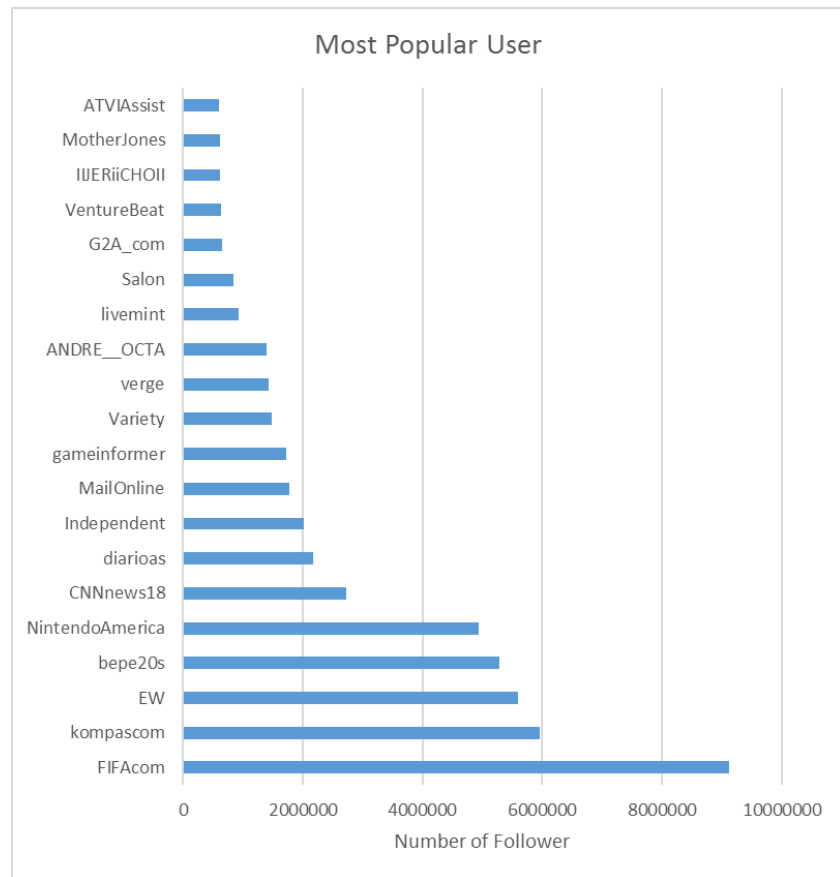
In addition, as can observe from the data of the game name, there is possibility that I expand my analysis to the popularity of game, or even platform. But since only incomplete game list are chosen, it might turn out to be biased. However, it need to be pointed out that with larger coverage of the topic and data, this is realizable.

3.3 How to Unify?

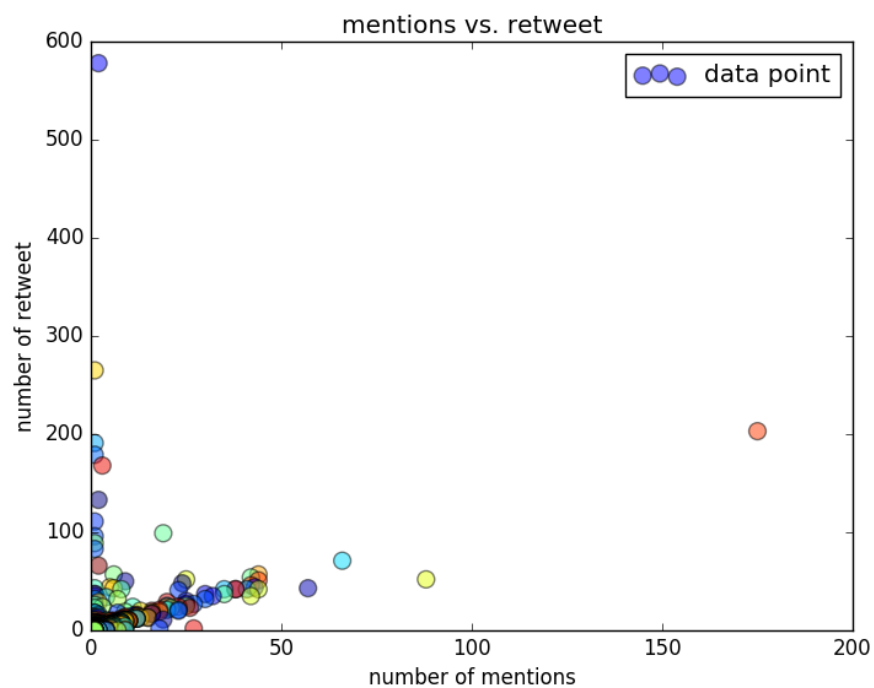
3.3.1 Twitter Metrics Unification

Metric 1 and 2 are easy to merge together as they are all sorted on number of followers. The merge process will force to retain the account that both publish and be mentioned under the same game topic, and is therefore a way to ensure user interactivity. The output file can be view at [*metric_one_two.csv*](#). The screenshot shows the top 20 entries, and I name it most popular involver.

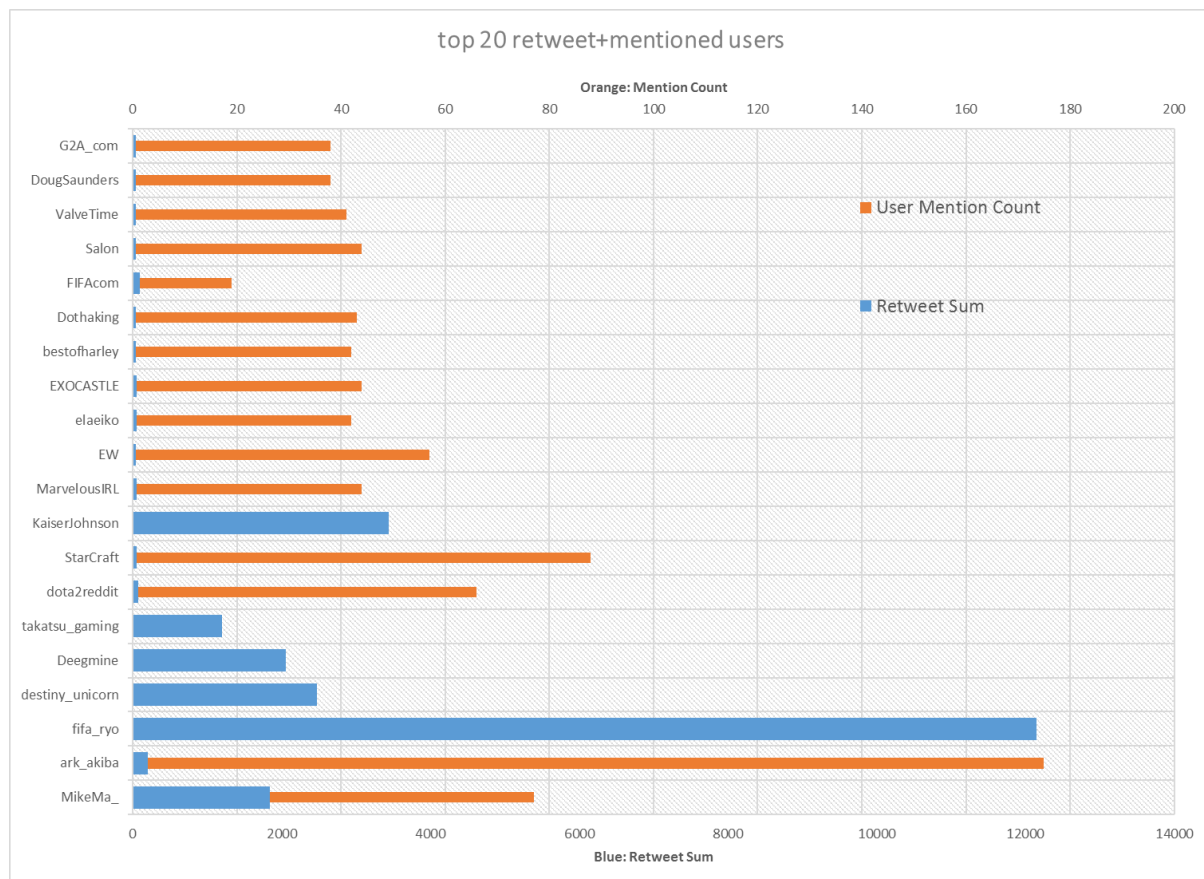
As can be observed from the diagram, previous entries like “YouTube” and “Rihanna” no longer appeared. This shows that these two accounts never publish game-related tweet under the topic they being mentioned in this data sample.



For Metric 3 and 4, I initiated another scatter plot to observe the relationship between mention and retweet:



Again, there is no obvious pattern to relate the two quantities. Therefore, to put the two parameters together, I calculated their product and sort according to it. And the following chart shows the top 20 when sorted by the product. The complete file is in [metric three four.csv](#).



I will cease my merge action to this stage, and leave 2 possible lists for Twitter. I could use product operation to join the 2 different results, but it does make any sense since the count for user being mentioned and sum of retweet also includes the effect of number of follower.

3.3.2 Overall Evaluation of both YouTube and Twitter

In total I generated 3 different outcomes for the top 100 influencers in gaming industry. 1 from YouTube and 2 and 3 from Twitter, which can be referred in files naming:

[output_youtube.csv](#)

[metric_one_two.csv](#)

[metric_three_four.csv](#)

They approached the problem in different ways, search for data in different database, and parameterized by different quantities, and are therefore different from each other. Typically for twitter platform, out of the top 100 records in each metrics, 28 are in common and is shown in the list below; and out of 1,000, 297 are in common.

	screen_name	Game_name
0	NintendoAmerica	Gaming
1	Independent	The Last of Us
2	Salon	The Last of Us
3	MyNintendoNews	Nintendo Land
4	DendiBoss	Dota2
5	StarCraft	StarCraft
6	EXOCASTLE	Candy Crush
7	TheObeyAlliance	Call of Duty
8	Sn_203069304	Dota2
9	Harada_TEKKEN	Battlefield
10	etserbu	Candy Crush
11	now7grandkids	The Last of Us
12	GIBiz	Grand Theft Auto
13	Nintenderos	Hyrule Warriors
14	NinEverything	Hyrule Warriors
15	DougSaunders	The Last of Us
16	dota2reddit	Dota2
17	Cs_Madrid	Gran Turismo
18	carnojoe	Uncharted
19	OverwatchFeed	Overwatch
20	PeerIGN	Nintendo Land
21	ColdplayAtlas	The Last of Us
22	ark_akiba	Counter Strike
23	ValveTime	Team Fortress
24	ShinersBR	Candy Crush
25	kanikahanda	The Last of Us
26	doope_jp	Uncharted
27	MikeMa_	Grand Theft Auto

This overlap suggested that the rationale behind the search is valid.

As a concluding remark, all 3 lists are generated with an emphasis on a particular quantity or platform, therefore, it is not possible to select and merge them into one single list. Because of their different focuses, they can be referred in different missions that would be discussed further in the next section.

4. Other Comments and Future Consideration

4.1 Choice of Platform

Being an amateur that had just recently been determined to build my career as a data analyst/scientist, I did not have a broad repository of skill to work across different types of software. Therefore, I choose Python which I had recently started to self-study in, as it is open-source, and provide many remarkable functionalities not only for general purpose but also particularly in data manipulation, where different extensions and libraries can be downloaded. At the same time. Its language syntax is also compact, making the coding process less tedious.

4.2 Different Top 100, Different Meanings

I finished my exploration with 3 different lists of names displayed. They can be used under different purposes.

List from the first two metrics is the absolute measurement of followers on Twitter. This list is more generalized in a sense that it has a particular emphasis on the absolute personal influence of the user to the public. The potential use for the list for example, game company can approach them to try out and publicize advertisement for the new release of game as to evoke public curiosity and awareness. This list irradiates to a broad range of audience not only confine to people who usually play games, hence to a degree more effective to attract **new player**.

List from the last two metrics is the relative popularity of user as both participants and authorities in a particular topic of game. This list is more confined in gaming industry and is more influential among heavy gamers. For propaganda purposes, this list could be used by game company that release new features to a particular game and help to **maintain the old players or intrigue them being more addictive/loyal** to the game.

Data from YouTube platform is a list of top subscribers, among them with a great expansion from commentary writer to professional gamer. This data has **a combination** of the two mentioned features. Audiences on YouTube not only approach these influencers for gaming advice, but also for entertainment purposes watching them making mockery of different games. Therefore they are suitable for both attracting new player and maintaining old ones, but the publication effect may only be restricted on YouTube.

4.3 Project Timing

As I was asking for an advancement of interviewing schedule and Mr. Scott Tend agreed to reschedule the technical phone interview to 9th-13th of September, the proposed duration for this project is 1 week (28th Aug – 4th Sep).

Day 1: Understand Python, preliminary search, timeline plan, study sample codes from Google

Day 2: Implementing API query, crawling data from YouTube, elementary analysis on samples

Day 3: Deciding on Parameters, design program flow, API query for Facebook (aborted)

Day 4: API query for Twitter, Elementary data analysis

Day 5: Run the query (6 hours), Design post-processing

Day 6: Optimization of Twitter parameters

Day 7: Graph generation, report writing

4.4 Future Work

- Build up **Dimension**: Include more Platforms, more Games, more Parameters, more Data
- More **Statistics**: the current data has no method to assess its credibility.
- Seek **Unification**: to find whether there is overlap across platforms/games, understand the differences and overlaps.
- Code **Optimization**

4.5 A Recap of the confounders and biases in the design/implementation of experiment

Some potential confounders/biases for the design of survey:

1. **Platform:** the social platform chosen may not be a good representative as the media for all the influencers to post. E.g. Stream
2. **Game:** Some influencer may not play any of the popular games listed in this search but would have large impact since they are famous for video commentary (therefore not restricted for one particular video).
3. **Incomplete Keyword Search:** influencer ID, alternative name for a game, byproduct of a game, other game communities, timing issue etc.
4. **Small Sampling Size/Data not representative**
5. **Incomplete Analysis (not including the activeness of user on Twitter)**

Appendix: Files included

Python script:

- *YouTube_final.py*
- *Twitter_one_keyword.py*
- *Twitter_two_postprocess.py*

Result files as csv:

- *Output_youtube.csv* (YouTube)
- *Metric_one_two.csv* (Twitter)
- *Metric_three_four.csv* (Twitter)

Intermediate file sample as csv:

- *All_sorted_follower.csv* (twitter metric 1 file)
- *All_sorted_follower_mention.csv* (twitter metric 2 file)
- *All_user_retweet.csv* (twitter metric 3 file)
- *All_user_mention.csv* (twitter metric 4 file)
- *Data_twitter_FIFA.csv* (twitter raw file)
- *Data_twitter_Pokemon Go.csv* (twitter raw file)