Date: 02/09/2016

Case Study 1

Names: Mohamad, Farah, Satish

Motivation about the, why the topic is interesting to you:

Data Collection:

The Data we collect from twitter is based on a search queries from tweets history.

Software used to code and collect data:

We used tweepy 3.5.0 search api to launch automated queries to gather daily samples from 02/01/2016 until 02/07/2016 that is related to death or people giving condolences about dead people. So, we structured our search queries in a way to sample 15000 tweet for every day of the given week for every given search key from the following list (**RIP, R.I.P, rest in peace, condolence**) and we were able to collect approximately 360000 tweets that worth more than 1.6 GB. The script ran over night, and took around 5 hours to complete, and stored in 4 different json files one for each search key that holds a week's worth of data around 90000 record each.

The Motivation behind the choosing the topic:

During the brainstorming period as a group we came up with multiple ideas that can be implemented for the given case study. We thought about politics, marketing, healthcare, sports... but finally a very interesting idea came up to our minds and we all felt very passionate about it. The idea initially started around a topic known as “Ghost Accounts on Social Media” and basically that topic covers mainly three types of accounts:

1- Inactive social media accounts for dead users

2- Inactive social media accounts created by machines

3- Inactive social media for users created and don't use it for more than 6 months

- Our Initial plan was to be able to get all the followers for a specific celebrity or popular twitter page and then do the analysis on the followers who are not active on the page for 6 months. Finding this set of people means they are ghost accounts for this specific page.
- The second step was out of the list of inactive users we found, we need to find how many of them are not active on there pages, in order to find out which set of them most likely to be machine created accounts or owners passed away.
- The third step is to do text mining on their recent activity and find out if there is any thing related to death mentioned by their friends and follower, which will indicate if they are most likely dead or not.

The Constraints and Restrictions:

Unfortunately, we got hit with many restrictions and limitations from twitter API platform, that prevented us from doing a lot of experiments we wished we could do. Because of the restrictions on users private info like (age, DOB, gender …), restrictions on retrieving user's activities like (last login, Favorites/Likes on tweets, comments mentioned on private pages …) and lastly but not least, the limitations twitter search api has on query history data. We hoped that we could query tweets/Activities on a given page from September which is 6 months ago, but Twitter only let you go back one week in your search and not more than that.

The deviation from the main idea:

Given these restrictions from Twitter API, we decided to proceed with our idea, but with reverse engineering the concept and making multiple assumptions, that will weaker the results, but still showing the intended analysis and the road map to do it. We will search for dead people based on people's tweets and classify them as ghost accounts.

How did you analyse the data?

We thought instead of finding the inactive followers of specific page, why we don’t search for tweets during the past week that talks about death of somebody and be able to narrow it as possible to figure out if this person is dead or not? So we sampled twitter

history from 02/01/2016 until 02/07/2016 and we retrieved 15000 tweets samples from each date for each search key from this set **(RIP, R.I.P, rest in peace, condolence**) which we thought it will be a good common keywords regarding death of someone.

After we managed to gather 360000 tweets, we noticed duplicates and retweets, so we need it to do something. Also we are interested in the people mentioned in the tweet "@Mentioned" which will be a bigger indicator on behalf of whom this death condolence issued. Also regarding @Mentioned we noticed many duplicates. Furthermore, we thought if we know the person who issues a death related tweet, that means he lost someone, so we could also provide analysis of people who lost someone during the last week on twitter.

Dealing with duplicates, we proceed all the tweets and extract from each tweet the list of screen names of the mentioned accounts, the screen names of the tweets creators, the screen names of the in_reply_to_screen_name in the tweets. We successfully managed to get the unique screen names for all the three categories. Then we noticed that some tweets could be a condolence between two people and the in_reply_to_screen_name is mentioned on the screen, for example, X Mom passed away and he tweeted about it, then Y is retweeting his tweet with condolences and mentioning him on the tweet. To avoid duplicates in our list of suspected dead people, we took the unique subset from @Mentioned lists that doesn't exist in the list contains people who lost someone.

Having the unique list of @Mentioned people suspected to be dead or ghost accounts, we went to check if there twitter profile is public then we can read their user_timeline and get the latest tweet/retweet done on the account (assumption made here, if the user didn't tweet or retweet after mentioned on death related tweet, it means for us he is inactive and could be dead, because we can't get the last login or Favorite/Like activity). We processed the list of suspected dead accounts and retrieve their last tweet/retweet on their account and compared the tweet date with the date of the death related tweet they mentioned on. If the last tweet/retweet date is greater than the date of the death related tweet he is mentioned on, it means he was active after it and he is most likely not dead, but the tweet is about somebody related to him or he knows, so we move him from the suspected dead list to lostSomeOne list. But if the date is smaller than the

death related tweet he will stay on the suspected list with more confidence that he could be dead.

What did you find in the data?

After Analyzing the 360000 tweets collect, below are some of the results we found:

Before Checking Last Activity of suspected dead account:

Number of unique accounts suspected dead = **743**

Number of unique accounts suspected lost someone = **2524**

Before Checking Last Activity of suspected dead account:

Number of unique accounts suspected dead = **3207**

Number of unique accounts suspected lost someone = **35**

Number of private accounts from suspected dead list = **14**

Table of suspected dead list after checking last activity:

Screan Name	Last Activity	Location	Dead (Y/N)
Sergioo_NikeAir	01/05/13 06:26 AM	Atlanta ✔	Y
Beyonce	08/19/13 08:31 PM		N
krisndrob	06/20/14 06:41 PM		N
Ryanschools	12/05/14 12:19 PM	India and UAE	N
LeMarquand	02/15/15 09:51 AM		Y
TrendLaRose_	07/20/15 03:29 AM		Y
davemirra	12/04/15 12:29 PM	Greenville, North Carolina	Y
UWC_IO	12/17/15 11:26 AM	Worldwide	N
itskekeloves__	12/19/15 11:00 PM	South CLT	Y
lawarehouselive	01/07/16 04:05 AM	Houston	Y
RyanIntlGrp	01/07/16 09:26 AM	India | UAE	N
GNev2	01/09/16 05:06 PM		N
ChinxMusic	01/19/16 12:53 AM	Queens, NY	N
Ventured_	01/24/16 11:05 PM	XO	N
eiyooyie	01/29/16 04:09 PM	Olongapo City, Philippines	N

ImranDhamrahPPP	01/30/16 08:42 AM	Sindh, Pakistan	N
quickmixx	01/30/16 05:36 PM	JAMAICA	N
hottunaband	01/30/16 07:22 PM	On the road...	N
SUGIZOofficial	01/31/16 06:23 AM	Tokyo	N

Real_Liam_Payne	02/01/16 12:46 AM	UK	N
BREEZYROASTS	02/01/16 11:14 AM	Manchester, England	N
boyd_chantelle	02/01/16 10:03 PM	Blackpool, England	N
PyongyangPimp	02/02/16 05:16 AM		N
OTY_Yungdex	02/02/16 06:20 AM		N
MamaGH	02/02/16 08:23 PM	earth!	N
SpeakMemoryOk	02/02/16 11:53 PM	Oklahoma City, OK	N
nyalan_jalan	02/03/16 03:58 AM	川崎あたり	N
Aidan2535	02/03/16 01:32 PM		N
SpencerDay	02/03/16 08:07	Los Angeles	N

	PM		
dfeingoldphoto	02/03/16 10:31 PM	NYC	N
i_m_p_e_r_i_u_m	02/04/16 10:38 PM		N
EarthWindFire	02/04/16 10:57 PM		N
MarthaPlimpton	02/04/16 11:02 PM	New York	N
QtipTheAbstract	02/04/16 11:41 PM	ubiquitous	N
reimansm	02/06/16 11:57 PM		N

After we manually analyzed the final suspected list of dead people, we confirmed the death of 6 people in our list. And the rest of the list are accounts mentioned on a death tweet who knows the person who died, so we are going to add them to the list of people lost someone which bring the final results to be like the following:

Number of people actually dead = **6**

Number of people lose someone = **3236**

Why the number of lost someone way bigger than the actual dead?

Basically, from the data we collected we noticed that many tweets mentions exactly the name of the person who died, but with mentioning his Twitter account. Our analysis is, old people have higher death rate than youth, and old people are less active on social media, so most likely they don't have twitter accounts.

Secondly, we found out that people uses death related words like RIP or rest in peace to make fun of certain people or topics, for example, #RIP_BlackBerry, RIP Donald Trumps….

We noticed that, the way someone will die will give bigger impact on the reaction of people on social media, for example "itskekeloves__", she passed away because of car accident, and she is young in her twenties, which made a lot of people sad and remembered her on social media. Another example, is "davemirra" he is a celebrity who suicide on Feb 04, and surprised a lot of his followers and fans, and also made them very active on twitter in remembering him and giving their condolences.

Frequency Analysis

Word Count:

Based on our topic to analyze on ghost account on twitter, We collected around 360000 tweets tagges with 4 key words we chose to filter with.

- R_I_P
- rest in peace
- RIP
- Condolence

statistics on word analysis:

Total Tweet Count : 360011

word analysis were implemented by getting all the unique words from all the tweets collected and filtering with the list of stopwords.

Top 5 word counts.

	word	Count
0	condolence	79478
1	prayers	13350
2	allah	13200
3	hang.	12900
4	strongly	12900
5	condemn	12900

Top 5 hashtags

0	Quetta	12900

1	QuettaBlast	12900
2	ALDUBBoojieWonderLand	9322
3	MarkFarren	3000
4	ALDUBPangalawangPagsubok	2850
5	Saudi	2204

Top 5 screen Names

0	TanzeelSHK	12900
1	me_gicana28	6636
2	derrycityfc	6450
3	aldenrichards02	2606
4	mainedcm	2606
5	Abunass3r	220

Popular tweet: popular tweet is determined by the tweet that is retweeted most number of times.

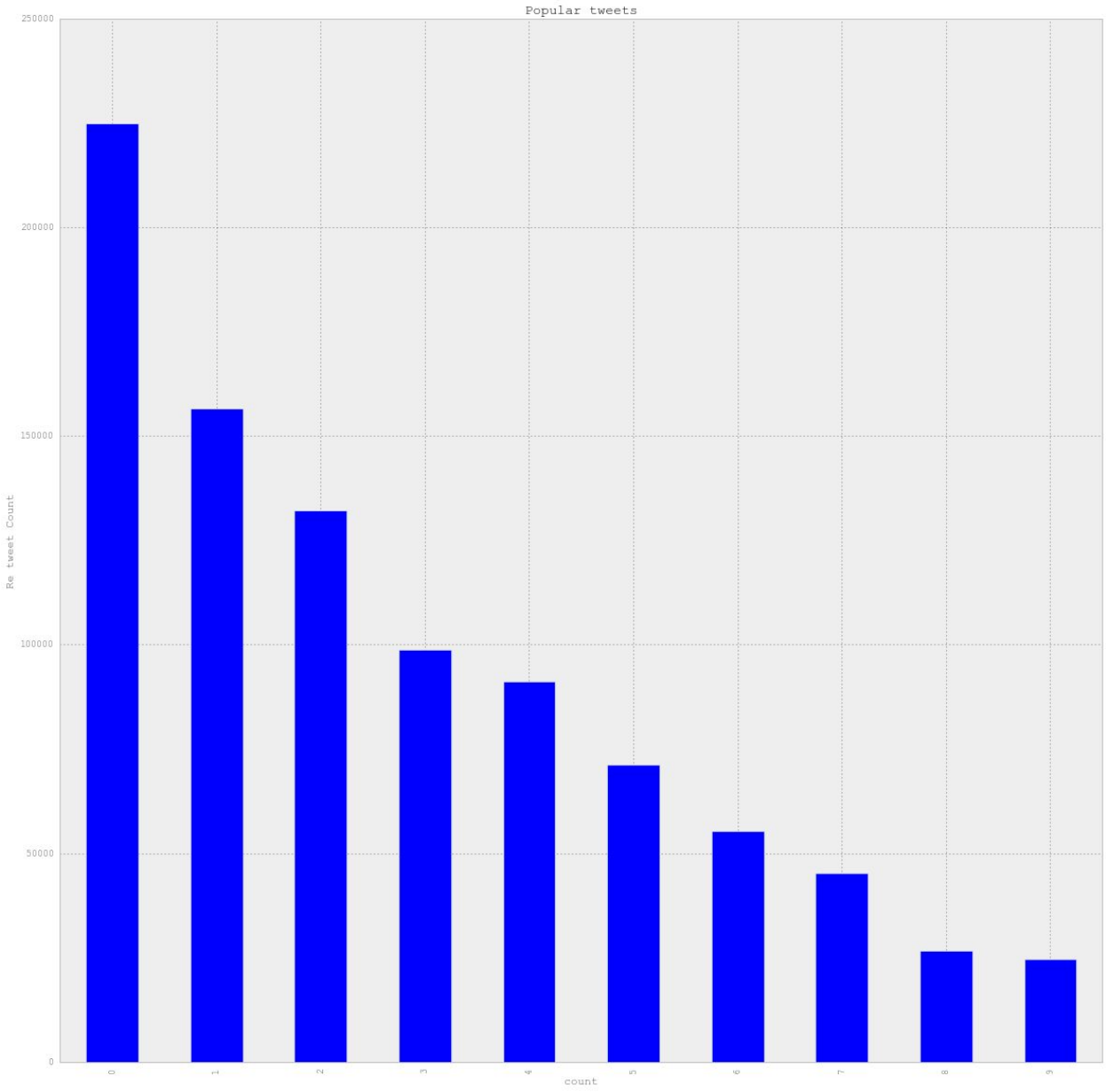
based on that definition, the **algorithm/pseudo** code for fetching the retweet is:

- parse every tweet to see if it has a "retweet" node
- Ignore the tweet that doesnt have "retweet" node 0 which meant not a RT.
- with the list of retweets only, parse each RT to get the retweet info and store it in to a dict.
- the dictioanry is created with key as the tweet ID.
- For every RT, get the retweet ID as key, User, create dt, text and retweet count as value.
- if RT id already exist in the dictionary, check the retweet count - if it > current retweet count, override the dictioanry with the new RT count.
- Now, we have a dictionary of keys with RT ids and vaues with RT count,user and text info.
- Plot a table with the content
- Plot a graph on top 10 retweet counts.

Most popular tweet:

499042364846268416	One of my all time favourites. Movies I grew up watching over and over again. A genius that will be so missed.\n\nR.I.P. Robin Williams.	224858	Harry_Styles

Top 10 retweet count graph:



Fetch the Friends and followers info.

1. We used the tweepy search API to get the friends and followers ID'd
2. used api user_lookup to get the used info based on the id's
3. obtained set of Followers and Friends list.
4. Get the common names that are in both the list to determine the mutual friends.

Conclusion:

Due to Twitter API limitation, we had to deviate a lot from our initial goal, but our passion about the idea of extracting death related information and being able to distinguish ghost accounts from active accounts, made us excited to proceed with the study and see what we can get. The fact that we were able to find dead person in our final result, indicates that our method can be enhanced for better analyzes and predictions. Ultimately, the ease in Twitter API with giving more information and allowing history search interval bigger than a week to be in months, made more account activity data available will help a lot in doing better analysis and coming up with better results.

Overall, it was a great project and we learned a lot about twitter api and python packages dealing with it. Learned and used AWS and GitHup with a great team spirit.

References:

Death and the Internet:

https://en.wikipedia.org/wiki/Death_and_the_Internet

Tweepy :

http://docs.tweepy.org/en/latest/api.html

Stopwords list:

https://github.com/ravikiranj/twitter-sentiment-analyzer/blob/master/data/feature_list/stopwords.txt

panda:

http://pandas.pydata.org/pandas-docs/stable/