Project Proposal

Team 5: Karlis Dagilis, Arielle Dupiton, Shashank Kava

Our idea is to make a twitter bot that will post tweets weekly - statistics together with programmed data visualizations on numbers, geolocation, common interests etc. about people who favorite or retweet fake news on twitter. We will also create and post statistics on how many times these people who are retweeting fake news twitter handles are retweeting current U.S. president.

To create such a twitter bot our stepwise approach would be as shared below:

a) Find a good and verified list of twitter handles that posts fake news online;
b) Verify ourselves top 10 fake news twitter handles;
c) Write a code that will extract twitter handle names from Twitter using the API of persons/bots who are retweeting tweets posted by handles from verified list;
d) Arrange extracted account names to a database - with columns like:
    - Initial tweet
    - Twitter handle name that posted the tweet
    - Twitter handle name of person who interacted with this tweet
    - Retweeted? (yes / no)
    - Favorited? (yes / no)
    - Coordinates of the location
    - Name of the place
    - How many times during past week this twitter account has retweeted current U.S. president
e) Manually or by writing a code drop off every account that might be bot itself or another fake news twitter handle (not a human);
f) Write a code that will make a top 10 lists on selected categories based on the statistics
g) Write a code that will make a bar charts and other data visualizations to highlight our automated weekly findings.

A sample of the extracted tweets with the location for the individual tweeting is shared below:

id	created_at	text	coordinates	place
5.6143E+17	1/31/2015 7:58	Because someone said, world is running out of Chocolate! #Foodgasm @ Chocolateria San Churro, Andheri http://t.co/9Qjb44ZOUa	[72.83161007, 19.13938922]}	Mumbai
5.6143E+17	1/31/2015 7:51	Life begins at the end of your comfort zone. #Travel #EveryPictureHasAStory @ Malvan http://t.co/zzoEWVOo2a	[73.4667, 16.0667]}	Malvan-Tarkarli
5.603E+17	1/28/2015 5:02	Perfect meal. @ Boulevard 9 Resort & Spa http://t.co/kiUFrcmf68	[72.83885768, 22.68764482]}	Nadiad
5.6009E+17	1/27/2015 15:04	Swaach Bharat! #YesWeCan	[72.8388719,	Nadiad

Challenges and limitations:

1. The coordinates and name of the place for individuals are not be available for multiple records in the dataset and hence cannot be shown in the visualization.
2. The location field had to be split in MS Excel as the attribute could not be retrieved using simple manipulation in Python.
3. The data collected will not be representative as we are not able to detect every twitter account that posts fake news on under their twitter handle;
4. There might be occasions when some tweets made by handle from our list are actually not posting fake news;
5. There might be some issues dividing who are real people who retweet and who are other fake news twitter handles or bots;
6. Most of these fake news sites have large following on Facebook by they don't have too many twitter followers or even don't have twitter accounts;
7. There might be occasions where some people retweet a fake news post to share it as a joke.

Future Scope:

1. If successful in writing the code, this can be also used to create other twitter bots that summarize stats if there is a special twitter handle list given.
2. The next stage can be detecting whether the tweet is actually a fake news before checking whether it is retweeted using machine learning instead of assuming that all the tweets by a Twitter handle are fake.

Word Count: 505