**Collecting Data from Twitter**

**DS 501 Case Study 1**

**Team #6 (Rahul Ghadge, Melanie Jutras, Praneeth Nooli, Debraj Roy, Sushma Tayanna)**

# Introduction

The purpose of this project was to gain an introduction to the field of data science by gathering, storing and analyzing sample twitter data. We were given the task of selecting a topic to study, searching tweets, analyzing tweets and tweet entities with frequency analysis, and getting all “friends” and “followers” of a popular twitter user. We were also assigned the task of coming up with a business question that twitter data could help to answer. The following sections of this document describe the process which was used, the data we collected and analysis of the data.

# Topic of Interest – Pokemon Go

Our team decided to focus our study on the topic of the popular location-based augmented reality game Pokemon Go. We decided on this topic for a number of reasons. First of all, we wanted to choose a topic with a likelihood of having many tweets associated with it. Additionally, due to the game being location-based and involving hundreds of different characters and millions of users across the globe, we felt that there would be a number of different analysis options and business questions we might be able to explore. Finally, we were looking for a fun topic that would be amusing and interesting to all team members and to the audience as well.

The Pokemon franchise started out as a video game created in 1995 for a handheld device called a Game Boy created by Nintendo. Although it was obviously not based upon GPS location or “augmented-reality”, the idea of the Pokemon creatures was very much the same as today’s game. The word Pokemon was derived from the Japanese brand name “Pocket Monsters”. Romanized and translated (**ポケ**ット**モン**スター ***Poke****tto****Mon****sutā*[**?**](https://en.wikipedia.org/wiki/Help:Installing_Japanese_character_sets)) (Source – Wikipedia)

The Pokemon Go application is free to download to mobile devices. The GPS location capabilities of the device are used to alert users to nearby virtual Pokemon creatures they can catch in the game. Users are encouraged to move from location to location in order to catch better, stronger, more diverse Pokemon for their collection. There are 151 Pokemon in the game and as of this date they are not all available to catch. There are many more features provided by the game, but for the purpose of presenting our data we have simplified the explanation to focus on this aspect of the users goal of collecting Pokemon.

# Data Collection Process

The data collection process for this project was done in phases as we were learning to work with new technology to be used for this course. Initially, data collection was done by using twitter’s polling API to get resulting tweets and save them to a file. As the course progressed, we were directed to use twitter’s streaming API and the popular free and open-source NoSQL database, MongoDB.

Although we collected tens of thousands of tweets over the course of this project, in the end our analysis presented here was performed on a much smaller data set of about 5000 objects. The lesson learned here was to be sure of the nature of the data that needs to be collected before getting too far into the project. Once data is consistently being gathered, proper database backup procedures must be in place. We had a corruption of the database late in the project and were unable to retrieve the bulk of the tweet data we had stored.

Some information about the data we collected is presented here. (See Graphs/Charts/Statistics of our data below) Our search criteria to retrieve streaming tweets was simply any tweet containing the keyword #PokemonGo.

The Top 10 Tweets collected, interestingly seem to be people using the PokemonGo hashtag in order to attract people to their sites.

+-------+--------------+----------------------------------------------------+

| Count | Screen Name | Text |

+-------+--------------+----------------------------------------------------+

| 38760 | PokemonGoApp | RT @PokemonGoApp: #PokemonGO Team Leaders |

| | | revealed! Team Valor: Candela. Team Mystic: |

| | | Blanche. Team Instinct: Spark. |

| | | <https://t.co/j05H5K0tfY> |

| 26316 | OMGitsAliA | RT @OMGitsAliA: I have a #PokemonGO PLUS to |

| | | giveaway to one of YOU! 😍🙌🏻 Just RETWEET + FOLLOW |

| | | for a chance to win. Good luck! 💙 https://t.c… |

| 26315 | OMGitsAliA | RT @OMGitsAliA: I have a #PokemonGO PLUS to |

| | | giveaway to one of YOU! 😍🙌🏻 Just RETWEET + FOLLOW |

| | | for a chance to win. Good luck! 💙 https://t.c… |

| 26313 | OMGitsAliA | RT @OMGitsAliA: I have a #PokemonGO PLUS to |

| | | giveaway to one of YOU! 😍🙌🏻 Just RETWEET + FOLLOW |

| | | for a chance to win. Good luck! 💙 https://t.c… |

| 26312 | OMGitsAliA | RT @OMGitsAliA: I have a #PokemonGO PLUS to |

| | | giveaway to one of YOU! 😍🙌🏻 Just RETWEET + FOLLOW |

| | | for a chance to win. Good luck! 💙 https://t.c… |

| 26311 | OMGitsAliA | RT @OMGitsAliA: I have a #PokemonGO PLUS to |

| | | giveaway to one of YOU! 😍🙌🏻 Just RETWEET + FOLLOW |

| | | for a chance to win. Good luck! 💙 https://t.c… |

| 26311 | OMGitsAliA | RT @OMGitsAliA: I have a #PokemonGO PLUS to |

| | | giveaway to one of YOU! 😍🙌🏻 Just RETWEET + FOLLOW |

| | | for a chance to win. Good luck! 💙 https://t.c… |

| 26310 | OMGitsAliA | RT @OMGitsAliA: I have a #PokemonGO PLUS to |

| | | giveaway to one of YOU! 😍🙌🏻 Just RETWEET + FOLLOW |

| | | for a chance to win. Good luck! 💙 https://t.c… |

| 26308 | OMGitsAliA | RT @OMGitsAliA: I have a #PokemonGO PLUS to |

| | | giveaway to one of YOU! 😍🙌🏻 Just RETWEET + FOLLOW |

| | | for a chance to win. Good luck! 💙 https://t.c… |

| 26307 | OMGitsAliA | RT @OMGitsAliA: I have a #PokemonGO PLUS to |

| | | giveaway to one of YOU! 😍🙌🏻 Just RETWEET + FOLLOW |

| | | for a chance to win. Good luck! 💙 https://t.c… |

|  |  |  |
| --- | --- | --- |
|  |  |  |

We analyzed tweets and tweet entities with frequency analysis. The top screen names are shown here.

+-----------------+-------+

| Screen Name | Count |

+-----------------+-------+

| OMGitsAliA | 1505 |

| PokemonGoMemes | 137 |

| tbg\_j | 78 |

| iGatoConBolas | 75 |

| surfacage | 60 |

| stickman8000 | 43 |

| james44ct357 | 34 |

| guardiacivil | 34 |

| pokemon\_player5 | 33 |

| PokemonGoApp | 22 |

| rocka374 | 22 |

| Newtiteuf | 21 |

| TGNSpanish | 19 |

| Oska09Beast | 18 |

| chrisvangelov | 17 |

| ivan\_pranjic97 | 15 |

| PoketweetUA | 15 |

| ShoutGamers | 13 |

| FameRTs | 13 |

| PokemonNewsHub | 12 |

| gaboooye | 11 |

| YTRetweets | 10 |

| YTRTs | 9 |

| ShoutRTs | 9 |

| Gamer\_RTweets | 9 |

| Wistitee | 9 |

| PokemonGoCorner | 9 |

| Pokemon\_Days | 9 |

| NAT\_Group | 8 |

| SuperBayleef | 8 |

+-----------------+-------+

The top hashtags found in tweets were as follows:

+------------+-------+

| Hashtag | Count |

+------------+-------+

| PokemonGO | 3066 |

| PokemonGo | 1368 |

| pokemongo | 514 |

| ポケモンGO | 235 |

| pokemon | 172 |

| Pokemon | 157 |

| POKEMONGO | 115 |

| FUNNY | 113 |

| game | 113 |

| europe | 113 |

| NEW | 113 |

| TURKEY | 113 |

| agario | 113 |

| TeamMystic | 88 |

| KW | 87 |

| minecraft | 79 |

| Pokecoins | 79 |

| illuminati | 69 |

| Ansbach | 57 |

| Pikachu | 50 |

| ポケモン | 48 |

| Squirtle | 48 |

| Abra | 46 |

| trolling | 44 |

| OC | 37 |

| Tustin | 36 |

| UW | 36 |

| Android | 35 |

| Alerta | 34 |

| Tordow | 34 |

+------------+-------+

The word frequency analysis was performed. However, our attempt to use the nltk toolkit in order to remove stop words from the top words was unsuccessful.

\*\*\* And here are the TOP 30 :

+--------------+-------+

| Words | Count |

+--------------+-------+

| to | 5125 |

| a | 3446 |

| #PokemonGO | 3038 |

| RT | 2277 |

| I | 1831 |

| for | 1719 |

| of | 1699 |

| have | 1652 |

| Just | 1636 |

| + | 1632 |

| one | 1630 |

| FOLLOW | 1629 |

| chance | 1620 |

| RETWEET | 1617 |

| giveaway | 1615 |

| PLUS | 1615 |

| YOU! | 1607 |

| win. | 1579 |

| Good | 1544 |

| luck! | 1498 |

| 💙 | 1456 |

| 😍🙌🏻 | 1451 |

| #PokemonGo | 1346 |

| https://t.c… | 1325 |

| @OMGitsAliA: | 1321 |

We were also asked to collect data on a popular twitter user. We chose to do a search using the screen\_name @JohnHanke . John Hanke is the CEO of Niantic labs, the company that makes the PokemonGo app. We fetched 327 total friends and 5000 followers. The following charts display friends and followers twitter ids and screen\_names.

+--------------------+--------------------+

| Friend ID | Friend Screen Name |

+--------------------+--------------------+

| 816653 | TechCrunch |

| 1636590253 | tim\_cook |

| 769080580302569473 | AgentKodama |

| 1724473160 | Skiplagged |

| 37666984 | nikkihaley |

| 105912240 | jimsteyer |

| 6331462 | jess |

| 1394399438 | JohnLegere |

| 16303106 | StephenAtHome |

| 20106796 | amberkanwar |

| 2227174303 | leanne\_graves |

| 81285400 | \_AlexHirsch |

| 2839430431 | PokemonGoApp |

| 21094279 | chillmage |

| 714975751 | TweetMcCool |

| 69489820 | erickasoTV |

| 14188906 | BryFitz |

| 96879107 | Pokemon |

| 19511752 | GabrielJR |

| 631577690 | NianticLabs |

|  |  |
| --- | --- |
|  |  |

+--------------------+----------------------+

| Follower ID | Follower Screen Name |

+--------------------+----------------------+

| 113594235 | eykinakamuta |

| 577964386 | yukarinmurasaki |

| 217434145 | ehdante |

| 4914835684 | AkaVevie |

| 759887823038214144 | Miguel69581435 |

| 4765106839 | valdenirluiz1 |

| 720515406532714496 | pokegotakatsuki |

| 778336178621800448 | argunbey1071 |

| 2164467097 | mikewsyd |

| 778358394063388677 | lucky\_lucario16 |

| 232204464 | gusgompy |

| 58844519 | prosimma |

| 1673113746 | garzarnancy |

| 2970697108 | Meeper12346 |

| 1211273371 | buddhablue21 |

| 778318908910477312 | Metro\_Instinct |

| 2204146394 | Domanated89 |

| 1391281 | the\_fresh\_one |

| 2726900475 | lhawkins90 |

| 2696492256 | ismar6300 |

+--------------------+----------------------+

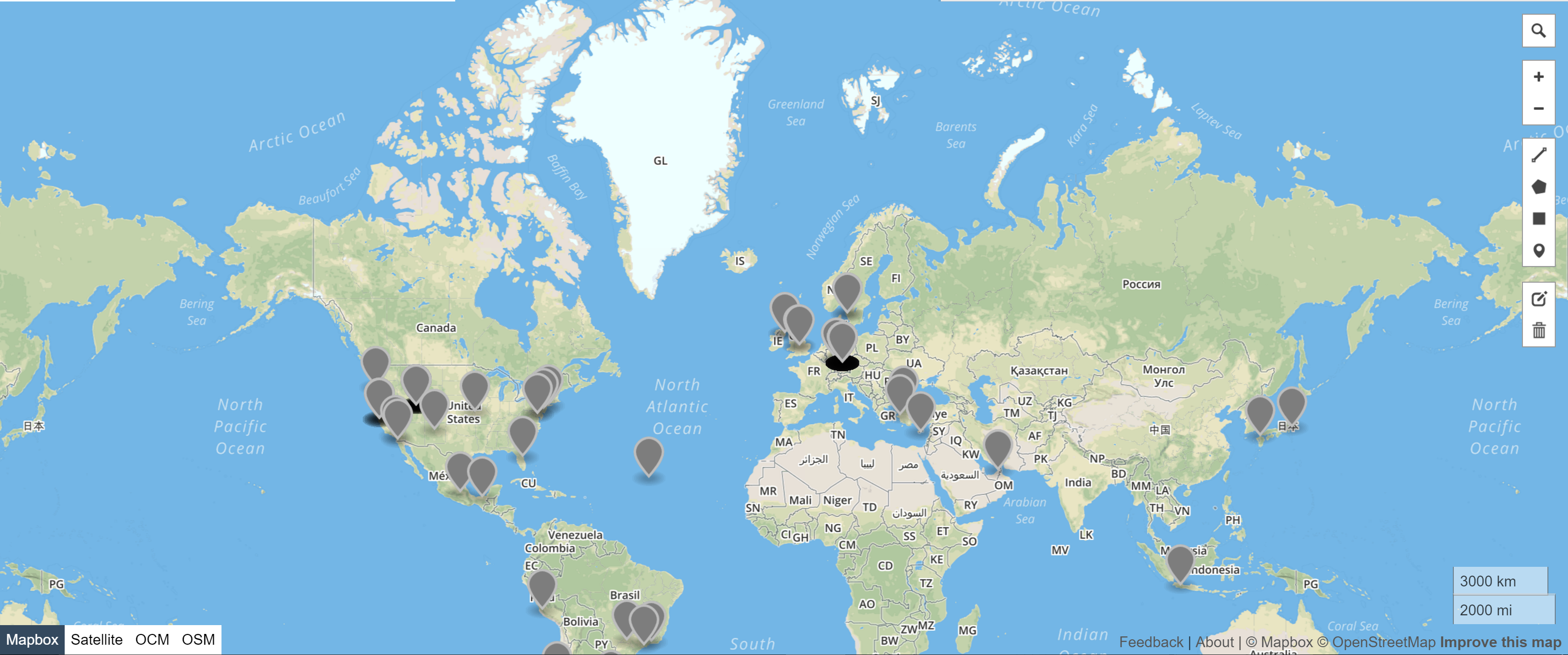
# Data Analysis

Analysis of the data was performed in order to understand what all of these #PokemonGo tweets were about! Early on in the process, we searched for a way to view the MongoDB data easily. We found a tool called Robomongo which is a Graphical User Interface alternative to the mongo shell. This was helpful for easy viewing of objects structures and their values.

In order to answer our business question and support the business use case, we were interested in analysis of geolocation of general tweets about the Pokemon Go application as well as specific tweets about the Pokemon characters as they relate to location. We quickly found that the tweet location is off by default and the user needs to opt-in to the service in order for that information to be stored. For that reason, it should be understood that the analysis done here as well as any analysis seen on twitter data with respect to geolocation will have this limitation or “bias” in the data that needs to be accounted for when drawing conclusions. We also briefly considered using the location field stored in the user’s profile, but that has its own complications because the user can enter whatever they like and the format of the data is unstructured.

We decided to use tweets that contained coordinates for our analysis. For this part of the analysis, we used instructions from Marco Bonzanini “Mining Twitter Data with Python” part 7 geolocation and interactive maps. The process involved gathering twitter data, looking for tweets where coordinates were given and writing them out to a GeoJSON data structure. This json file was then able to be used as input to another tool called geojson.io which plotted the points on a world map which can be seen below.

<http://bl.ocks.org/d/1921b74a3dad5410200097440ae7bb2b>



Next, we repeated this process for tweets about each of the top five most popular Pokemon characters and plotted those on a map as well.

Note – these maps are actually interactive. When given the url, a user can click on tweet object on the map to see the tweet at that location. This was a very exciting discovery and a powerful way to be able to visualize twitter data.

# Business Use Case

In this Case Study we were assigned the task of coming up with a business question that twitter data could help to answer. We decided to explore the question “Can the content and location of user’s tweets about the PokemonGo application provide useful information for future business decisions?”

The business case involves finding the location of where people are tweeting about PokemonGo as well as whether or not tweets about particular pokemon characters are somehow tied to location. If it can be determined that a location has very few tweets, the company may do an analysis of where pokestops exist or which pokemon characters might appeal to a particular location to try and attract more users. (Pokestops are popular gathering points for users because they are able to gather items needed for the game)

The idea is that some characters may be more popular in particular regions than others. For example, in Massachusetts, people may love the “Tauros” whereas in California they may love the “Jynx”. If there is a connection between location and twitter character then the business could use this information to make decision about which new characters should be included in future releases. Perhaps, certain characters would be more popular in some parts of the world as opposed to others.

Twitter data can help a company in many ways. Tweet location and sentiment analysis are two pieces of information that are becoming widely available with the advances in technology and data science.

# Future Scope

As discussed, geo location is turned off by default for twitter users. In order to draw conclusions based on location one would need to compile a considerable amount of data for analysis. Additionally, user sentiment is an area of great importance when analyzing tweets. The number of times someone tweets about a topic means nothing if we do not know whether the tweets are positive or negative. Identify popularity of each pokemon character along with user sentiment in various geo locations in order to determine whether specific new characters should be released only in certain geographic locations in the future.

During the course of this project we also considered the study of how a separate business entity might profit from tweets about Pokemon Go. The analysis we did could be used by a separate business entity to somehow attract game players to their location.

# Conclusion

In conclusion, we have gathered, analyzed and summarized data for thousands of tweets about the topic of the popular mobile application PokemonGo. With millions of players all over the world, there is great potential to determine future marketing plans and budgeting of resources. With the explosion of the internet of things, this data will continue to increase in volume and provide invaluable business insight.