Analyzing Twitter Data for #DataScience

by

Kevin Drever, Ryan Sawasaki, Lester Yang

in the

Graduate Division

of the

University of California, Berkeley

datascience@berkeley

Abstract

Analyzing Twitter Data for #DataScience

by

Kevin Drever, Ryan Sawasaki, Lester Yang

in Data Science

University of California, Berkeley

,

With over 500 million tweets every day, Twitter is arguably the leader in disseminating current events and capturing users' reactions, perceptions, and opinions to these occurrences. The events along with the user responses, gives Twitter a tremendous impact in shaping the worldwide conversation on topics of interest.

As soon as Twitter, along with similar social media, became this critical piece of global culture, decision-makers desired to understand and monitor the sphere, but they have struggled. The problem facing these decision-makers is deriving functional information from leading sources of relevant data with high volume and velocity but comprised of unstructured text. The purpose of this project is to use Twitter and the topic of #DataScience to break down unstructured user inputs to produce insightful and actionable information.

This project used a sample of Tweets utilizing #DataScience. The analysis focused on breaking out thought leaders in, and related topics to #DataScience. In addition, the project determined geographic locations of activity and public sentiment within the hashtag. The results showed Twitter is a valuable source of information critical in evaluating current trends in society.

What does Twitter reveal about Data Science from November 10th 2019 to November 25th 2019? The answer is much.

Contents

Chapter 1

Project Framework

1.1 The Focus

Twitter maintains Application Programming Interfaces (APIs), allowing developers access to Twitter data. Twitter structures its data in JavaScript Object Notation or JSON. JSON supports a structure comparable to nested dictionaries in Python. For each Tweet, there exists a Tweet object that contains four other JSON objects. The four JSON objects are the User object, the Entities object, the Extended Entities object, and the Place object. Each of these objects contains a nested dictionary with key, value pairs of data.

The User object contains various information about the Twitter user. Much of the most valuable information for data analysis is found in this object. Please see the appendix for the complete listing of the data available in the user object, but a few noteworthy fields found in this object include the user identification information, the user counting information (follow count, like count, etc.), and user text fields.

The Entities object contains added metadata and context information not found in the User object. Information related to hashtags, polls, links, and other interactive media is in this object. The Extended Entities object contains additional datastores for media that may become necessary in some instances. An example is attaching multiple photographs in a Tweet. No patterns of the Extended Entities object are contained in the appendix as the object is flexible based on the type of media, and a neat rendering is not possible. This object was mostly unused in this project as it did not contain data useful in hashtag characterization.

The final object is the Places object. This object contains geographic information, including geo-tags. Please see the appendix for further details.

1.2 Data Cleaning

The project used the Tweepy package and Python programming language to access the Twitter API. The Tweepy package provided easy to use methods for obtaining various JSON Twitter objects. The `.Cursor()` method allows for the retrieval of a hashtag, and in this

project, the `Tweepy.Cursor()` method accessed all the necessary #DataScience information. Other Tweepy methods used include the `Tweepy.get_user()` method which returned information regarding a specific user and the `Tweepy.lookup_users()` method. After accessing the information using the Tweepy Python package via the Twitter API, the data required formatting using Pythons JSON package. The JSON package formatted the data, so it became ready for storage in a Pandas data frame. Once stored in Pandas, the Tweeter data was sufficiently clean for exporting to a CSV file for permanent storage. The CSV file allowed further distribution of the data as well as importation into specific python scripts used to derive insight during the execution of this project. Additional data cleaning techniques are specified in the individual results sections as needed.

1.3 Data Limitations

In this section we list some potential problems and caution the readers of this report should keep in mind.

- This study does not incorporate statistical methods, and as a result, the insights gained pertain solely to the sample data and cannot describe the population.
- Objects relating to geo-data, coordinates, and places are part of developer accounts premium plan, unretrievable without a monetary subscription. Without funding, this project does not include insights derived from the premium data sources.
- Without the Premium Twitter subscription, the volume and duration of retrievable Tweets are limited. With the volume and data range limitations, the sample size was limited.
- The project did not use machine learning techniques on data per project instructions, even when machine learning produces optimal insight.

Chapter 2

Influential Users

Twitter is a popular community for data science professionals. From podcasts to presentations at conference panels, You can follow me on Twitter, and My Twitter handle is... seems to be the common last phrase of professionals. As such, in this section of the report, the aim is to learn about the biggest Data Science influencers and the contents of their Tweets.

2.1 Users

First, the project Identified the most influential users in the sample data. Prominence is a difficult concept to operationalize since it suggests the user is possibly changing the future trajectory of data science or motivating new or curious discussions in the community. Unfortunately, the use of time-series techniques did not provide descriptive information because of the datas limited collection duration. Fortunately, Twitter is a social media platform that naturally captures effect in one-dimensional metrics such as $favorites$ and $retweets$. For a single user, the users impact on the community can be measured with the accounts $followers\_count$ and $listed\_ount$, and a single tweets effect is measured with $favoritesandretweets$. Like any social media platform, followers/friends do not accurately measure how well the users voice is heard and agreed/disagreed-with on the platform due to inactive accounts. For simplicity, let this project calculated influence with two methods:

- rf_ratio: Dividing the average number of retweets a user received by the number of followers the same user has.
- ff_ratio: Dividing the average number of favorites a user received by the number of followers the same user has.

If every follower retweeted or favorited every tweet, the user would have a perfect score of 1. Furthermore, if the tweets captured in the sample data set is trending, a user can receive a one as well. These two new metrics provide a crude method of measuring a users influence while accounting for inactive followers. Between these two metrics, we consider retweets more important, as it requires substantial active effort for other users to retweet.

In comparison, when a user favorites a Tweet, the action is simply one click. Graphed in Figure 2.1 and Figure 2.2 is the top 50 followed users in our data set and their corresponding influence ratios.

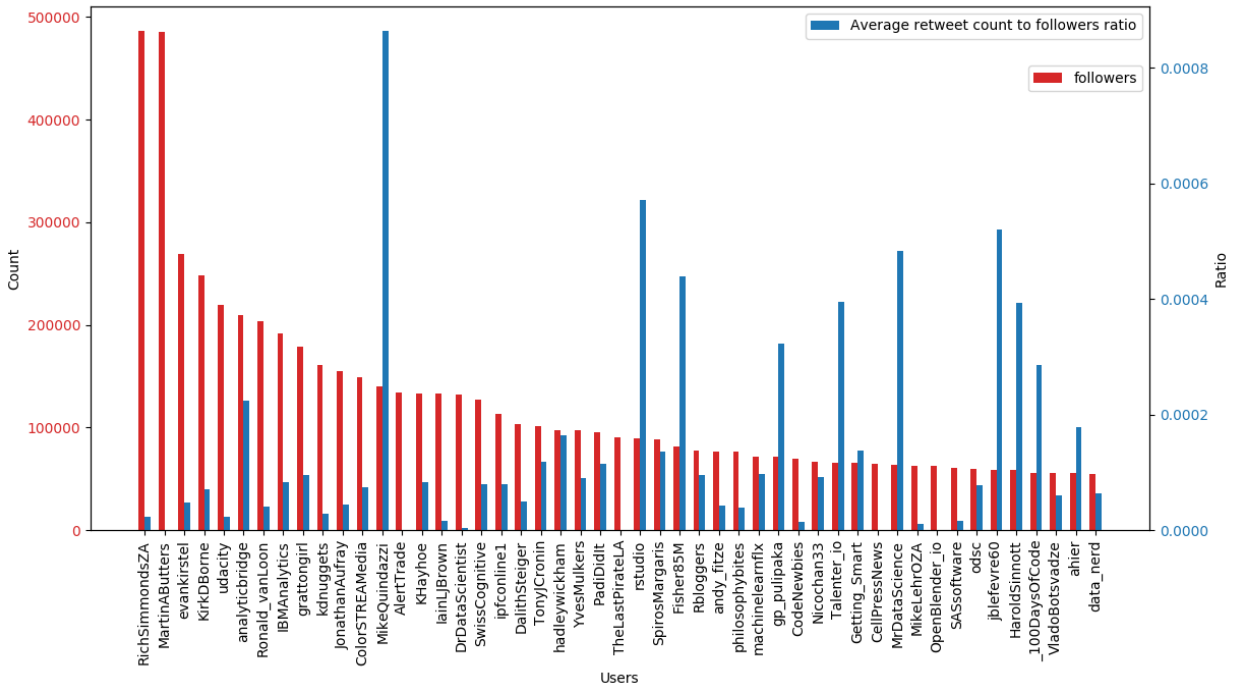

Figure 2.1: Top 50 users by followers and corresponding rf_ratio.

The data showed that the top 2 followed users are `@RichSimmondsZA` and `@MartinAButters`, each with nearly half a million followers. This number is two hundred thousand followers ahead of the third most followed user. However, both their rf_ratio and ff_ratio are extremely low relative to their follower's count. If we look at these accounts, both `@RichSimmondsZA` and `@MartinAButters` are marketing professionals with a dubious number of followers, copious number of tweets, and little to no retweet/favorites on most of the tweets. On the other hand, we notice from Figure 2.1 the users with high rf ratio relative to their follower count include `@analyticbridge`, `@MikeQuindazzi`, `@hadeleywickham`, `@rstudio`, `@Fisher85M`, `@gp_pulipaka`, and a few more. All these users have a significant number of followers near 100,000. Analyzing the data from their accounts one by one showed that they have average retweet and favorite counts hovering from 10 to 100 for most of their posts. The users on occasion post trending Tweets with over several hundred retweets and favorite counts. Figure 2.1 and Figure 2.2 seems to tell the same story, possibly suggesting a significant correlation between rf_ratio and f_ratio, ie, $retweets$ and $favorites$. Additionally, we see that RStudio is a favorite tool for data professionals, with the highest rf_ratio and second-highest ff_ratio amongst the top, followed accounts.

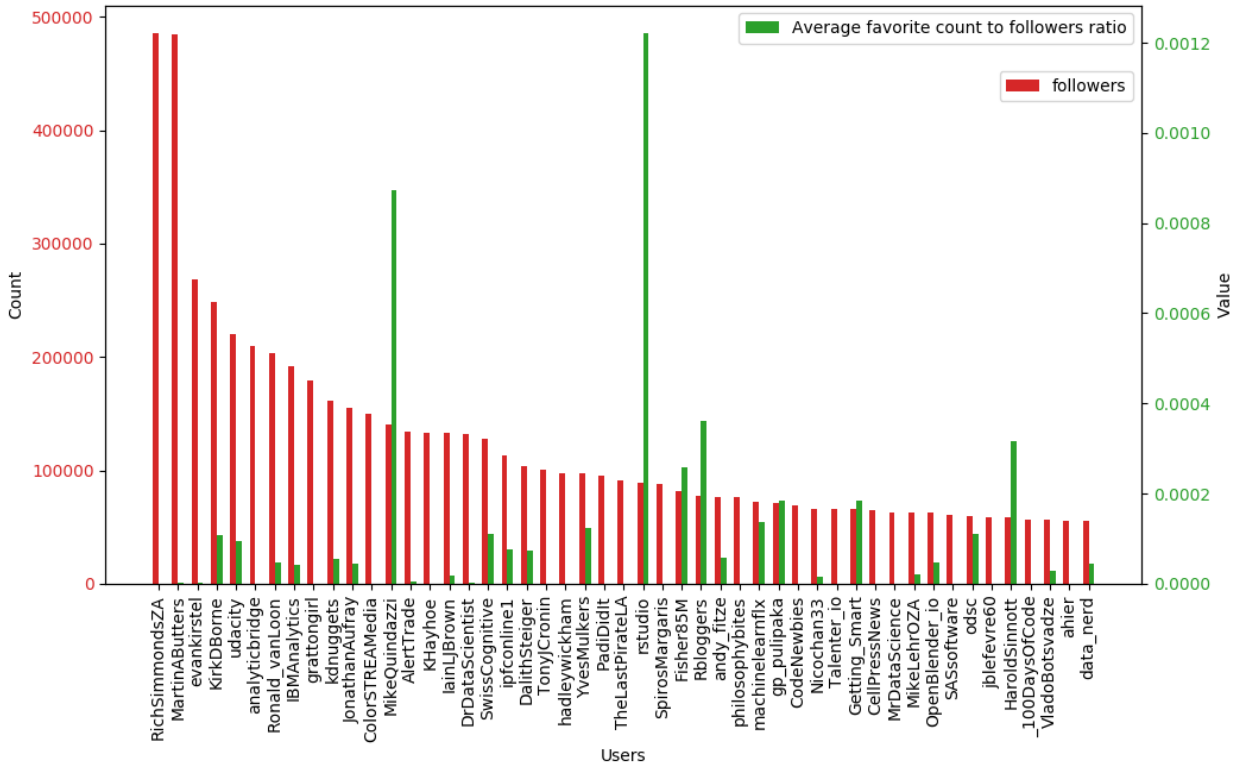

Figure 2.2: Top 50 users by followers and corresponding ff_ratio.

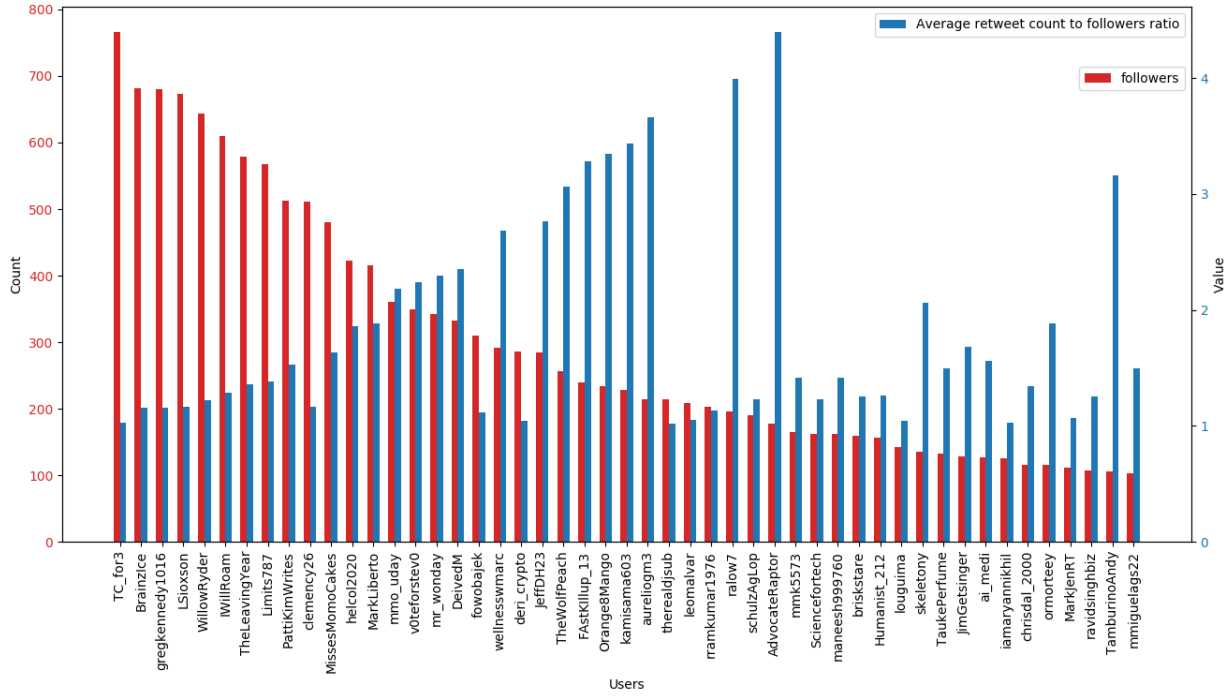

Figure 2.3: Top 50 users by followers with $rf\_ratio \geq 1$.

The project also looked at the trending tweets with particularly high rf_ratio values that are equal to or above 1. Out of the 9000 rows of tweets we have collected, only 203 rows coming from 176 unique users have $rf\_ratio \geq 1$, and Figure 2.3 shows the top 50 followed users. However, the top 50 most followed subset in this group with the highest rf_ratio values is very different from expected. Most of these accounts have nothing to do with data science or data. Instead, they appear to be a collection of random users.

When the word frequency of the self-description field is plotted, as shown in Figure 2.4, we see that besides the expected data, developer, scientist, keywords related to the 2020 Presidential candidate Andrew Yang also appeared frequently. As such, the accounts of users with the higher rf_ratio values, such as `AdvocateRaptor`, `ralow7`, `aureliogm3`, appear to be Andrew Yang supporters. We suspect that the time frame for which we made the Twitter API call for happens to contain a period of time during which the Twitter community supporting Andrew Yang tweeted about #DataScience. More investigation will be required to confirm this hypothesis.

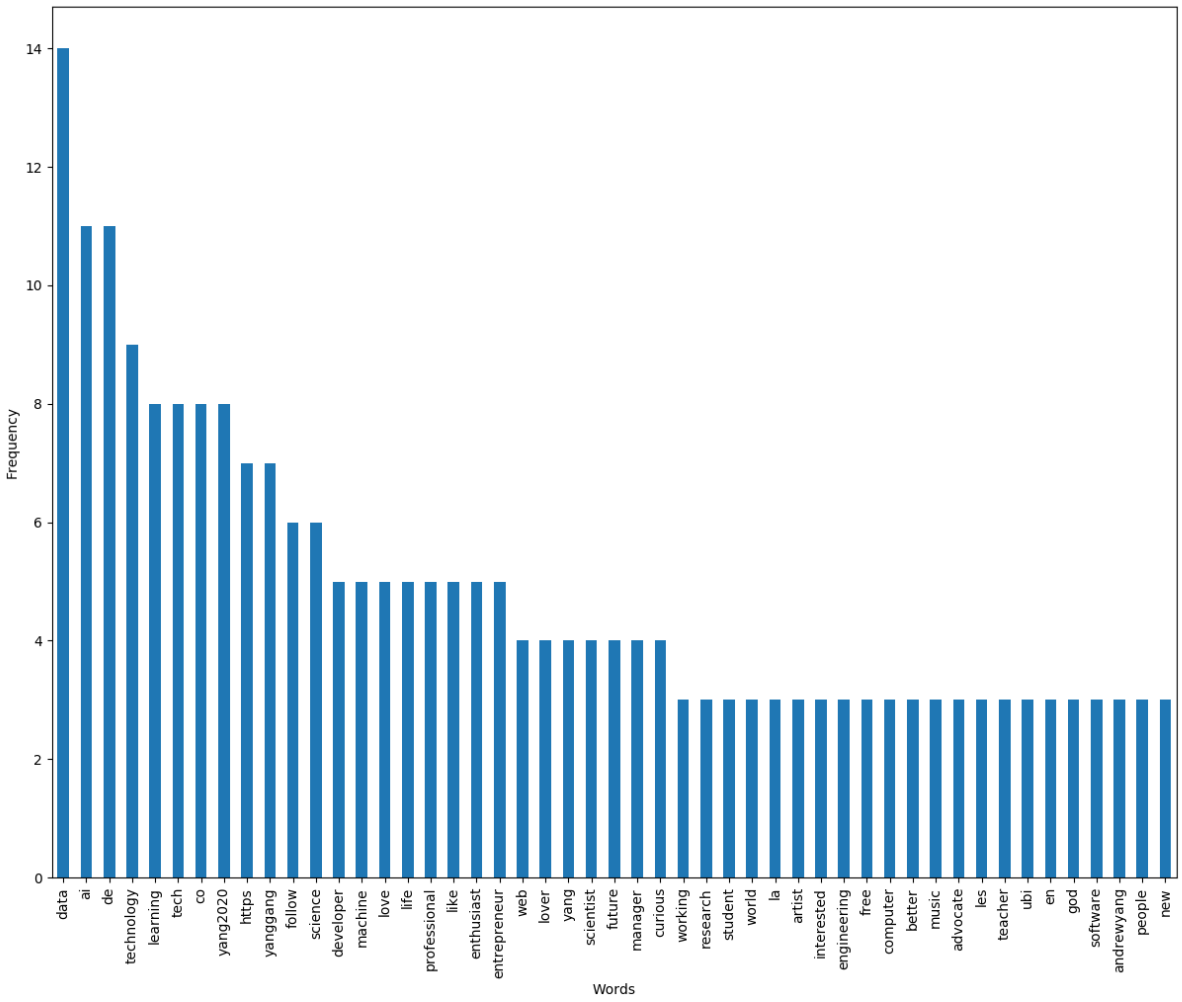

Figure 2.4: Top 50 words high rf_ratio value users used in the self-description fields.

Compared to Figure 2.4, Figure 2.5 generated the word frequency count for all 9000 rows of tweets. Judging from the top 50 words used in self-descriptions, notice a few key points:

- Business oriented community seems to dominate with keywords such as *analytics*, *business*, *founder*, and *marketing* ranking higher than what we would consider as practitioner keywords such as *research*, *developer*, *phd*, *engineer*, and *scientist*. Taking this point into account, it might be easy to understand why some of the most used words are buzz words such as *ai*, *bigdata*, or *blockchain*.

- A large number of users likes to link a url link in their description, signified by keywords such as *https* and *co*.

- In terms of occupation, the top self-described occupation in rank order appears to be: scientist, founder, developer, phd, engineer, director, entrepreneur, student, consultant, and author.

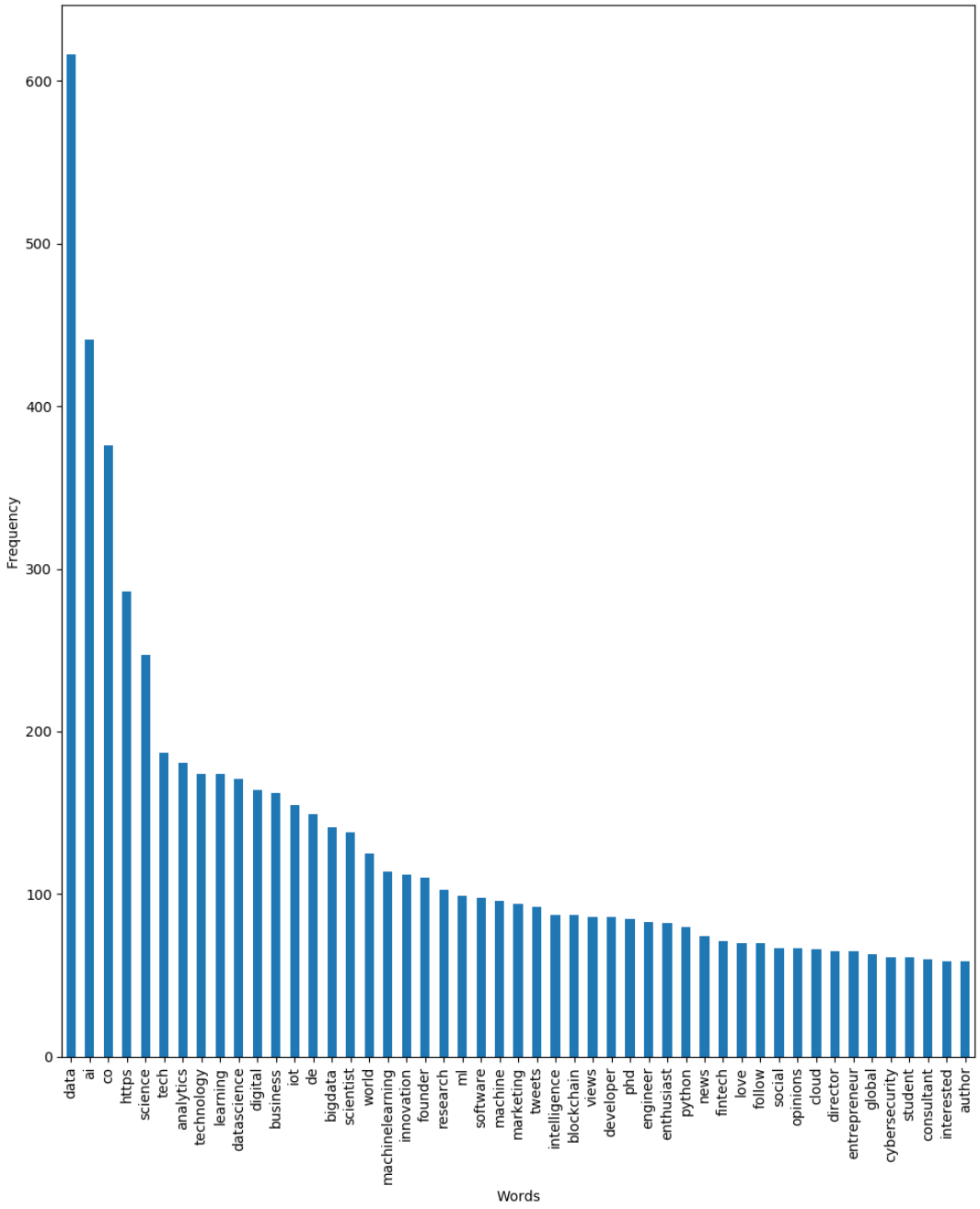

Figure 2.5: Top 50 words users used in the self-description fields.

It is also interesting to identify the Twitter user's device. The *source* field returned by the API captures both the domain and name of the source from which the tweets originate. Figure 2.6 shows that most tweets in the 9000 rows are made from the browser version of Twitter, mobile tweeting from Android phones. Apple phones come next in that order, and then If This Then That (IFTTT) ranks in third. Also, notice that Hootsuite Inc. listed in the 6th place contributed a significant amount of tweets. Since Hootsuite is a social media marketing platform, the data suggests these tweets either contain marketing content or aim to promote certain products. This was expected, as discussed in the introduction.

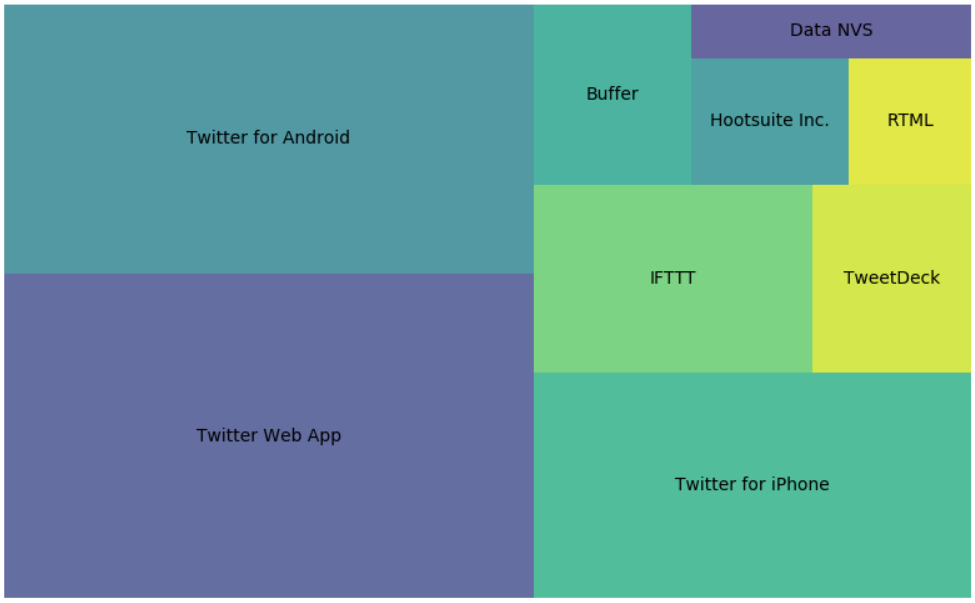

Figure 2.6: Top 10 sources from which users made the tweets.

2.2 What are they saying

Just as important as identifying te influential users, it is essential to know what these users are talking about. In Figure 2.7, we have plot the word frequency of all 9000 tweets, after removing common English stopwords with NLTK. As expected, buzz words dominated the top ranks, Twitter is a social media platform afterall. What is more interesting is the specific types of technologies and the user handles that are mentioned. The twitter handles mentioned are a proxy for the number of retweets a user receives. Based on the plot, these popular users are `@gp_pulipaka`, `KirkDBorne`, `MikeQuindazzi`, and `RichardEudes`.

The specific technologies that are popular are Python, AMP, PyTorch, R, JavaScript, and ReactJS.

Interestingly, the data suggests categorize of users displayed in Figure 2.1 The three main groups are influencers, practitioners(data scientists, CDO, PhD, etc.), and organizations. Attempts can be made to learn more regarding the difference between each of these groups by quarrying more data on them. The influencers that were selected are MikeQuindazzi, Fisher85M, HaroldSinnott, jblefevre60. The practitioners are KirkDBorne, gp_pulipaka, MrDataScience, hadleywickham, and the organizations are analyticbridge, Talenter_io, rstudio, and _100DaysOfCode. The selection method here is prone to bias because the data were chosen from some of the users from Figure 2.1 with relatively high rf_ratio. Therefore they are certainly not representative of their corresponding groups in the sample data set.

Additionally, the project quarried the Twitter API with built-in search operators that removes retweets from the search result. For influencers, 238 rows of data returned when quarrying for 1000 data points, suggesting that the operator has removed 762 retweets. For practitioners, 547 rows returned, and for organizations 217 rows returned. The word frequencies are summarized in Table 2.1. For influencers, it appears that tagging either themselves, or other users appears to be very common, as well as the expected buzz words. Meanwhile, for practitioners, mostly technologies and jargon were mentioned. Lastly, we see that organizations used a mix of buzz words and words such as *join*, *us* presumably for recruiting/advertising purposes.

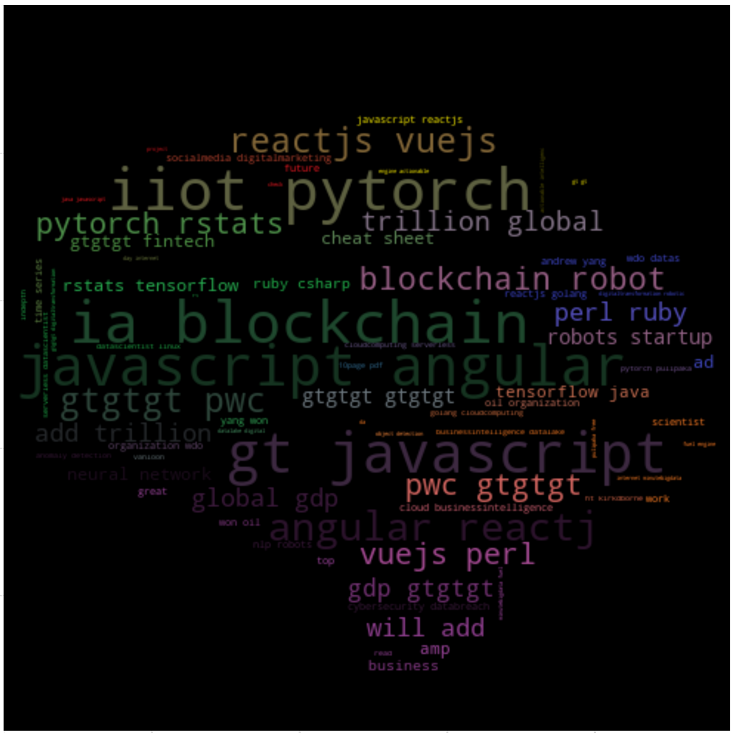

Figure 2.8: Word cloud of words used by #DataScience Twitter users in sample data.

Influencer		Practitioner		Organization	
gt	732	https	999	https	249
https	271	co	999	co	249
co	271	datascience	443	data	69
mikequindazzi	125	bigdata	429	learning	33
via	123	ai	411	machine	25
ai	119	machinelearning	394	via	25
iot	114	python	269	wnxsdqrbfm	24
fisher85m	74	iot	247	ai	23
spirosmargaris	61	rstats	245	science	23
evankirstel	57	analytics	245	help	18
robotics	57	datascientist	242	big	14
paula_piccard	49	iiot	241	2	12
andi_staub	49	javascript	240	ethereum	11
ipfconline1	49	serverless	239	blockchain	11
kalydeoo	47	cloudcomputing	239	using	11
ym78200	47	golang	238	4	10
sebbourguignon	46	linux	238	3	10
labordeolivier	45	reactjs	238	join	10
bigdata	45	tensorflow	197	us	10
haroldsinnott	45	pytorch	192	intelligence	9
diioannid	44	java	176	r	9
machinelearning	40	deeplearning	128	analytics	9
richsimmondsza	38	abdsc	85	get	9
fintech	37	statistics	79	python	9
4ir	37	datascientists	74	connect	9
mallys_	37	data	72	official	9
jblefevre60	35	algorithms	71	link	8
hitpol	35	mathematics	52	artificial	8
futureofwork	34	books	48	scientist	8

Table 2.1: Top words used and corresponding frequency by the 3 categorizes of #DataScience Twitter users.

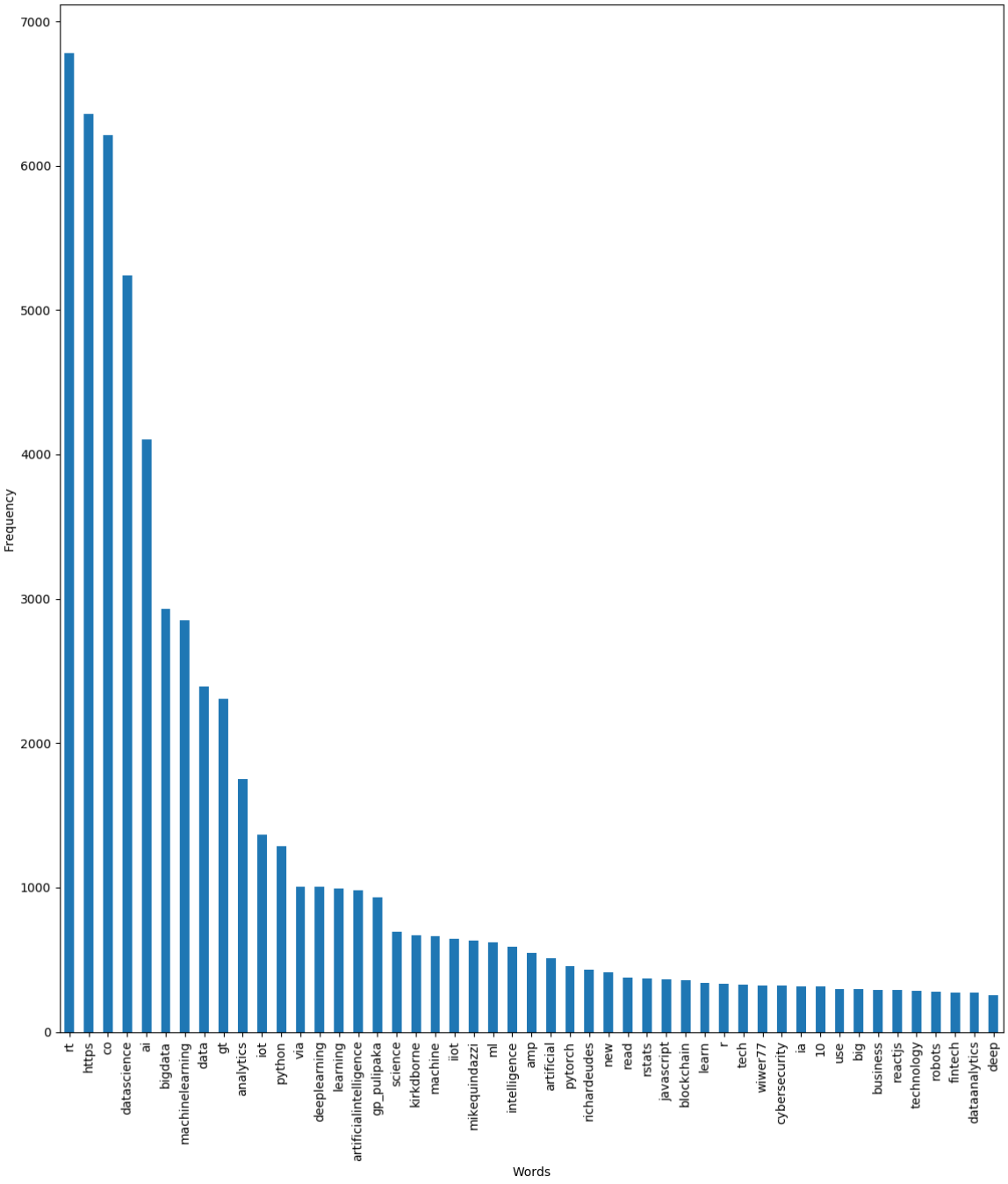

Figure 2.7: Top 50 words users used in the tweets.

Chapter 3

Location Analysis

To gain a better understanding of the reach and spread of data science throughout the world, this project analyzed the location data of twitter users that use the #DataScience. There was the expectation that the results would show that the majority of users are located in traditional tech-based locations. However, the analysis provided insight into other countries and cities that may have a growing data science industry.

Before beginning the analysis, the location data needed to be cleaned. The location data is input by the user, which created an added challenge for preparing the data for analysis. Most of the information scraped from twitter contained multiple tweets from the same users. To avoid double-counting locations duplicate screen names were removed. The filtering reduced the 9000 data points down to approximately 3000. Places with a single occurrence were also removed. Most of these data points were not identifiable locations, such as emojis and the location: Global. After removing the NaN data, the remaining data totaled approximately 1300 locations.

3.1 US Cities Analysis

U.S. locations were analyzed first. To pull the U.S. locations from the 1300 locations, identification and characterization of how most American users input their location were necessary. The majority of U.S. locations are identified by city-state, but there are also cases of state-USA, city-USA, and city only. Using a list of American state abbreviations, data was pulled into a data frame and split by the city-state combination. Additionally, city only values were added by matching the towns that were already in the data frame. Using this data to populate a data frame of U.S. cities produced the following plot.

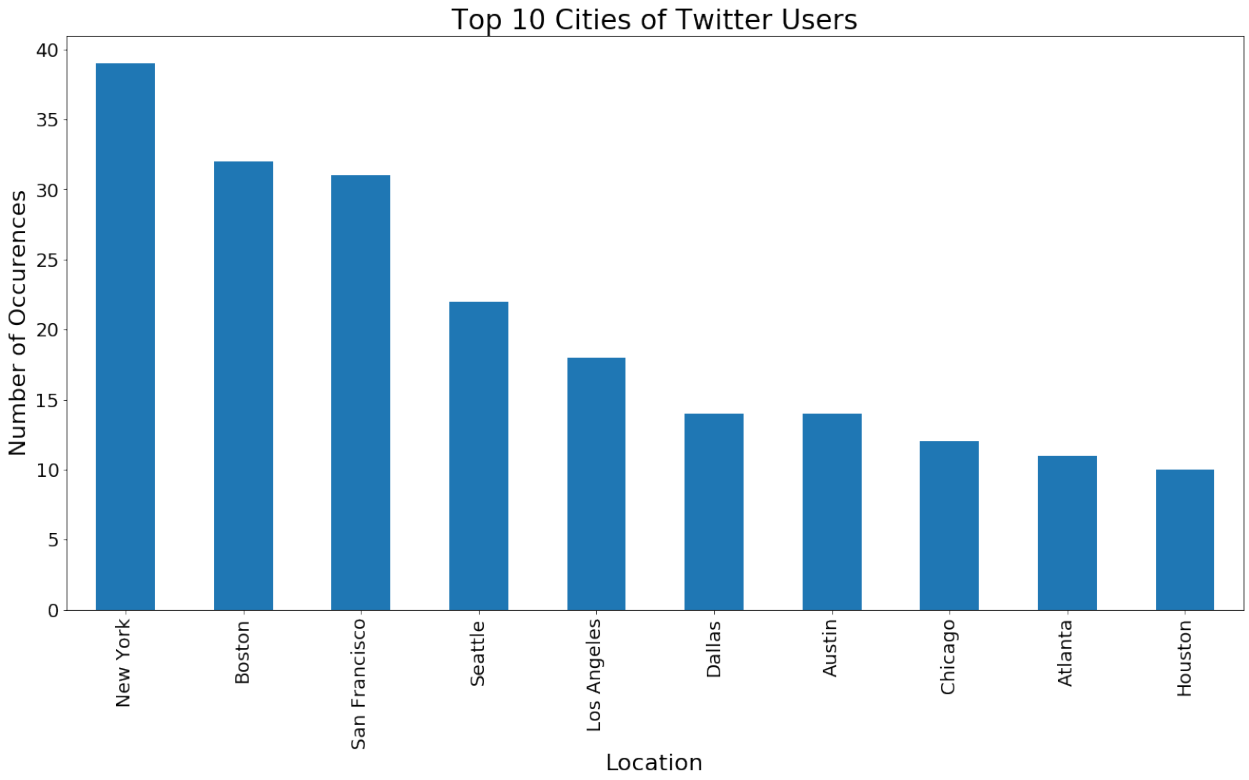

Figure 3.1: Top 10 cities of Twitter users.

As expected, many of the top U.S. locations are from major tech hubs. It should be noted that San Francisco does not encompass the broader Bay Area. Neighboring cities such as San Jose and Oakland were kept separate. While Austin has been steadily growing into a significant-tech city, the data suggests that the entire state of Texas may also be growing in the data science field, as Dallas and Houston have also joined the list of top cities in this analysis.

3.2 US States Analysis

Using the same method of pulling the city-state combinations, aggregation of the data was possible, and a data frame of the most common states was created. The data was read into the states.shp file, which contained U.S. state information, including geometry properties for mapping. Combining the two data frames and utilizing geopandas to assist in working with the geospatial data, produced the following choropleth map:

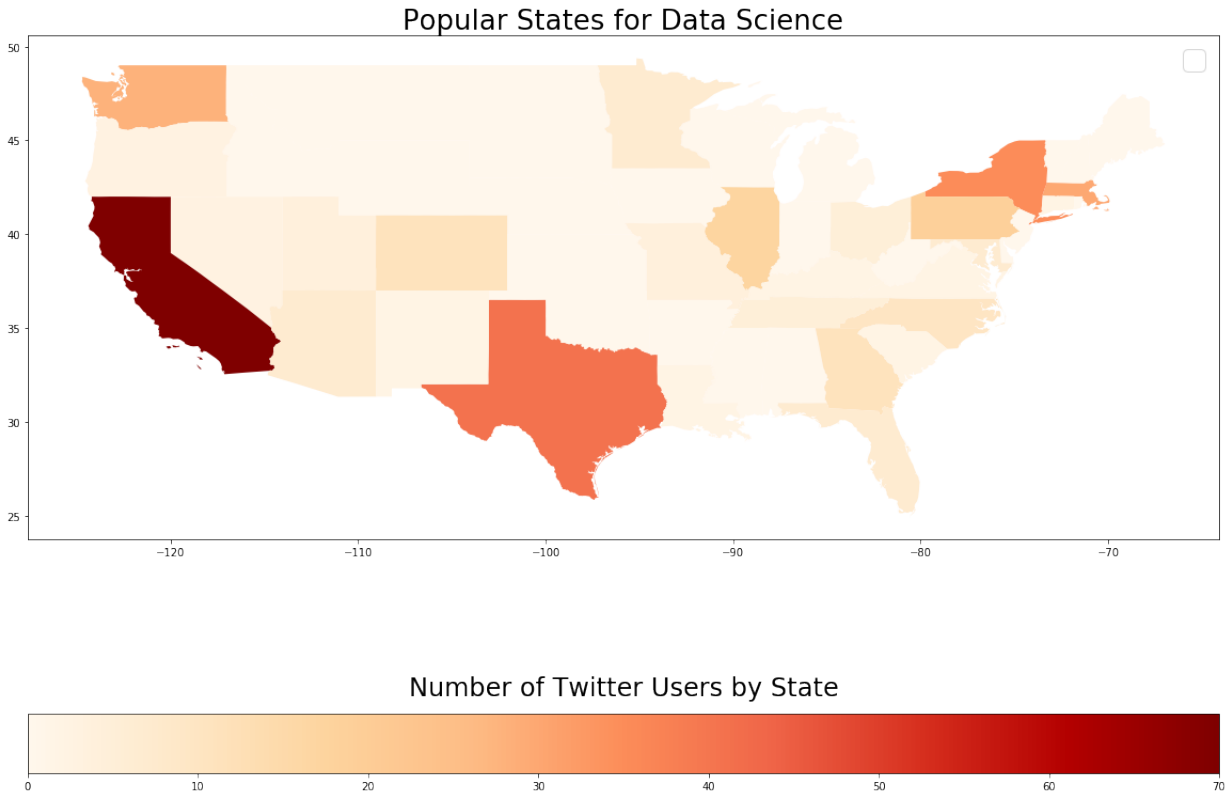

Figure 3.2: Twitter user heat map.

This visualization is consistent with the findings from the US city analysis. Combining the other Bay Area cities with San Francisco, and also including Los Angeles, pushed California to the top of the state list. Also, combining all of the Texas cities shows that Texas is in line with New York when it comes to the popularity of data science.

3.3 Country Analysis

Next, Non-U.S. countries' data were analyzed. The data cleaning for Non-US countries was challenging; given the multiple ways, people identify their location. For example, people in England note their location as England, London, UK, United Kingdom, and various combinations of the four. After wrangling the data, a world map showing the top international locations of Twitter users was produced.

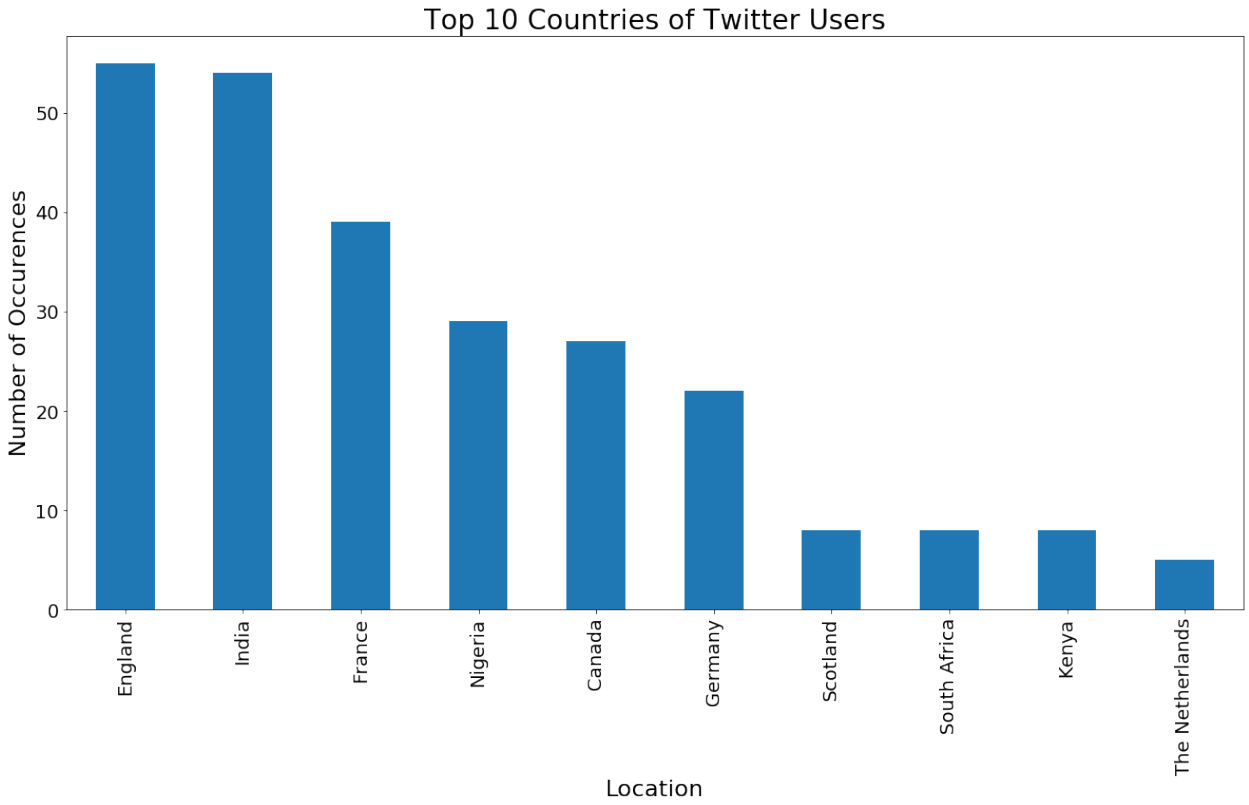

Figure 3.3: Top 10 countries of Twitter users.

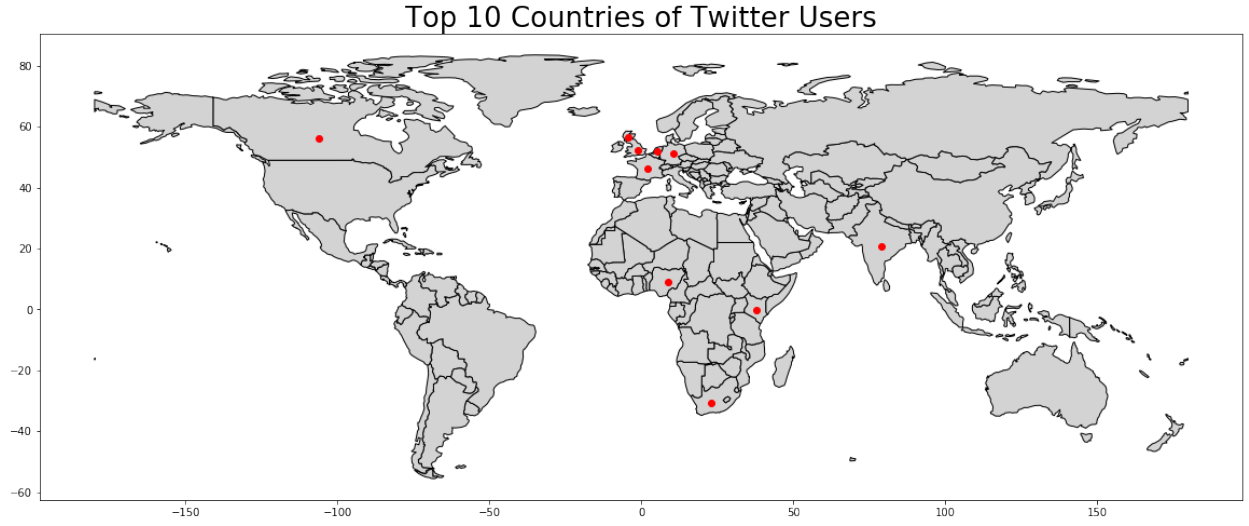

Figure 3.4: Top 10 countries of Twitter users.

This visualization shows that most twitter users using the #DataScience are concentrated in Europe. However, of particular interest is a large number of people located in Africa, which is traditionally not considered a tech center. Many people are using the #DataScience situated in Nigeria, which placed in the top 5 countries. Also, Kenya and South Africa set reasonably high. While there may be several reasons for the high number of people interested in data science in Nigeria, it would be interesting to investigate the data further to see if Africa is an emerging tech market, or if another explanation.

Chapter 4

Sentiment Analysis

Sentiment analysis is an essential technique for analyzing social media user's general perception on a given topic. There exist many methods to calculate sentiment, many of which utilize machine learning algorithms that are out of the scope of this report. Instead of machine learning, this project focused on non-machine learning algorithms that still provide impactful insights. This section describes two sentiment analysis techniques used to describe Twitter data.

4.1 Basic Bag of Words

This project used a bag of words counting method to determine a sentiment score for a given Tweet. The algorithm compared each word of the 9000 tweets against two lists. One list contained approximately 3,000 positive words, while the other list contained nearly 5,500 negative words. Each instance of a positive word located in a given tweet generated a +1 score, while each negative word produced a -1 score. After code execution, 622 tweets scored less than zero and therefore had a negative sentiment. Five thousand three hundred thirty-five tweets scored more than zero having a positive view. , tweets had no score, and were neutral. The highest scoring tweet was found at index 8412 and described the Top 7 data science use cases in trust and security. The lowest scoring tweet speaks to the dangers of data science; batch daemon spawn agi takeover deepfake deluge bias crisis how scared should you be?

count	mean	std	min	max
9000.0	0.927222	1.233912	-4	10

Table 4.1

The data shows a slightly positive sentiment overall. The results could be used to infer a generally positive perception of data science among Twitter users.

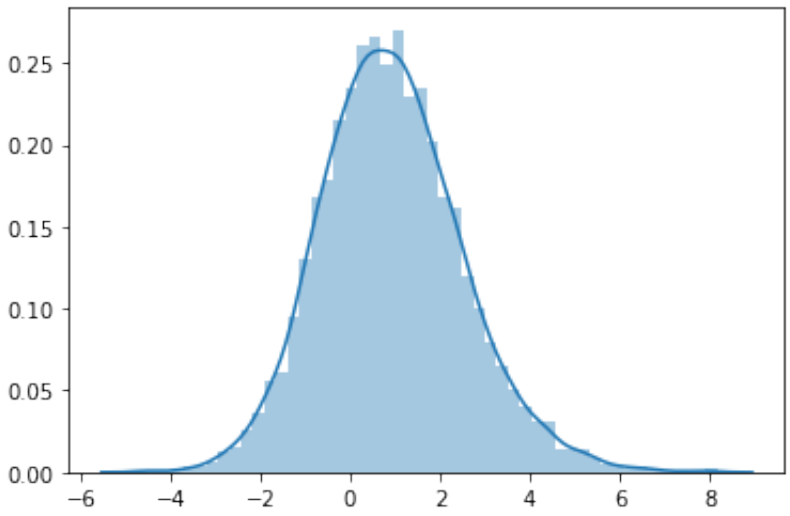

Figure 4.1: Diagnostic plots for sentiment analysis.

4.2 Semantic Orientation Method

Peter Turney developed the second Sentiment Analysis technique used during this project and detailed it in his paper *Thumbs Up or Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.* Peter Turney used the semantic orientation of phrases to generate a general thumbs up (positive) or thumbs down (negative) review. He developed the PMI or Pointwise Mutual Information (PMI) algorithm. The algorithm he invented found reference words (adjective or adverbs) in a given piece of text and compared them with the words excellent or poor. A review and scoring of the algorithm found it accurate between $66 - 84\%$ of the time based on the subject matter.

This project modified the Turney algorithm to focus on positive and negative sentiment in topics. The algorithm compared two adjacent words for the sentiment. If the 1st word of the two had a positive connotation, then the second word increased in the score. For example, outstanding presentation scores positive while boring presentation scores negatively for the word presentation. The algorithm determined the final score by comparing the number of instances the noun word occurred within the 9000 tweets as well as how many times the sentiment or descriptor word was positive or negative.

Insight derived from the analysis shows how Twitter users view certain aspects of Data Science. For example, some of the most negatively charged words in this projects data set were AI, Big Data, Data Cleaning, privatizing genomic, edits, data breach, and bias, listed in no specific order. While positive words also listed in no particular order built, experience, career, business, DataViz, hackathon, aistrategy, 100daysofcode, and developers.

Top Positive Words			Top Negative Words		
Rank	Word	Score	Rank	Word	Score
1	build	134.2308745	1	aggravatin	-43.92198465
2	experience	105.5655669	2	bias	-39.7915167
3	areas	104.6125472	3	agi	-39.48457934
4	career	102.4074867	4	foradversarial	-35.97839044
5	business	100.3723485	5	broadly	-35.97839044
6	bigdataanalytics	99.38668603	6	datacleaning	-34.85285956
7	article	97.74805827	7	aiwritten	-34.68293456
8	customer	91.97451605	8	eat	-34.68293456
9	check	88.15806773	9	cook	-34.68293456
10	cancer	78.43440056	10	humans	-34.55597036
11	aistrategy	75.36694073	11	privatizinggenomics	-34.18989455
12	hackathon	74.29365241	12	duty	-32.26789706
13	code	73.58655124	13	heighten	-32.18043422
14	100daysofcode	72.58978615	14	convenient	-31.19877197
15	insight	72.51501931	15	batch	-31.02267047

Table 4.2: Top positive and negative words used by #DataScience Twitter users in sample data.

The positive and negatively correlated terms are useful to businesses, researchers, and those who participate in data science. Seeing that 100daysofcode scored high in positive sentiment suggests users generally favor or talk about the coding challenge. Similarly, the hackathon also obtained a positive score. Hackathon is a collaborative conference for software developers and programmers. It may not be a surprise to a data scientist that terms like data cleaning received a negative sentiment score. Below is a list of the top positive and negatives ratings after the removal of stop words and phrases with little meaning, such as numbers and dates.

.1 Appendix

Attribute	Type	Description
id	Int64	The integer representation of the unique identifier for this User. This number is greater than 53 bits, and some programming languages may have difficulty/silent defects in interpreting it. Using a signed 64 bit integer for storing this identifier is safe. Use id_str for fetching the identifier to stay on the safe side. See Twitter IDs, JSON and Snowflake . Example: "id": 6253282
id_str	String	The string representation of the unique identifier for this User. Implementations should use this rather than the large, possibly unconsumable integer in id. Example: "id_str": "6253282"
name	String	The name of the user, as theyve defined it. Not necessarily a persons name. Typically capped at 50 characters, but subject to change. Example: "name": "Twitter API"
screen_name	String	The screen name, handle, or alias that this user identifies themselves with. screen_names are unique but subject to change. Use id_str as a user identifier whenever possible. Typically a maximum of 15 characters long, but some historical accounts may exist with longer names. Example: "screen_name": "twitterapi"
location	String	Nullable . The user-defined location for this accounts profile. Not necessarily a location, nor machine-parseable. This field will occasionally be fuzzily interpreted by the Search service. Example: "location": "San Francisco, CA"

Table .3: User data dictionary

Attribute	Type	Description
derived	Arrays of Enrichment Objects	Enterprise APIs only Collection of Enrichment metadata derived for user. Provides the Profile Geo Enrichment metadata. See referenced documentation for more information, including JSON data dictionaries. Example: "derived":"locations": ["country":"United States","country_code":"US", "locality":"Denver"]
url	String	Nullable . A URL provided by the user in association with their profile. Example: "url": "https://developer.twitter.com"
description	String	Nullable . The user-defined UTF-8 string describing their account. Example: "description": "The Real Twitter API."
protected	Boolean	When true, indicates that this user has chosen to protect their Tweets. See About Public and Protected Tweets . Example: "protected": true
verified	Boolean	When true, indicates that the user has a verified account. See Verified Accounts . Example: "verified": false
followers_count	Int	The number of followers this account currently has. Under certain conditions of duress, this field will temporarily indicate 0. Example: "followers_count": 21
friends_count	Int	The number of users this account is following (AKA their followings). Under certain conditions of duress, this field will temporarily indicate 0. Example: "friends_count": 32

Table .4: User data dictionary (cont.)

Attribute	Type	Description
listed_count	Int	The number of public lists that this user is a member of. Example: "listed_count": 9274
favourites_count	Int	The number of Tweets this user has liked in the accounts lifetime. British spelling used in the field name for historical reasons. Example: "favourites_count": 13
statuses_count	Int	The number of Tweets (including retweets) issued by the user. Example: "statuses_count": 42
created_at	String	The UTC datetime that the user account was created on Twitter. Example: "created_at": "Mon Nov 29 21:18:15 +0000 2010"
profile_banner_url	String	The HTTPS-based URL pointing to the standard web representation of the users uploaded profile banner. By adding a final path element of the URL, it is possible to obtain different image sizes optimized for specific displays. For size variants, please see User Profile Images and Banners. Example: "profile_banner_url": "https://si0.twimg.com/profile_banners/819797/1348102824"

Table .5: User data dictionary (cont.)

Attribute	Type	Description
profile_image _url_https	String	A HTTPS-based URL pointing to the users profile image. Example: "profile_image_url_https": "https://abs.twimg.com/sticky/default _profile_images/default_profile_normal.png"
default_profile	Boolean	When true, indicates that the user has not altered the theme or background of their user profile. Example: "default_profile": false
default_profile _image	Boolean	When true, indicates that the user has not uploaded their own profile image and a default image is used instead. Example: "default_profile_image": false
withheld_in _countries	Array of String	When present, indicates a list of uppercase two-letter country codes this content is withheld from. Twitter supports the following non-country values for this field: XX - Content is withheld in all countries XY - Content is withheld due to a DMCA request. Example: "withheld_in_countries": ["GR", "HK", "MY"]
withheld_scope	String	When present, indicates that the content being withheld is a user. Example: "withheld_scope": "user

Table .6: User data dictionary (cont.)