datascience@berkeley

Analyzing Twitter Data for #DataScience

Kevin Drever / Ryan Sawasaki / Lester Yang
W200 Python Fundamentals for Data Science

Outline

- Data Acquisition
- Influential Users
- Location Analysis
- Sentiment Analysis

#

Problem Statement: Twitter data contains valuable insight into current trends, perceptions, and opinions but decision-makers struggle to derive insight from the high volume, high velocity, unstructured text.

Research Question: How can Twitter #DataScience be characterized?

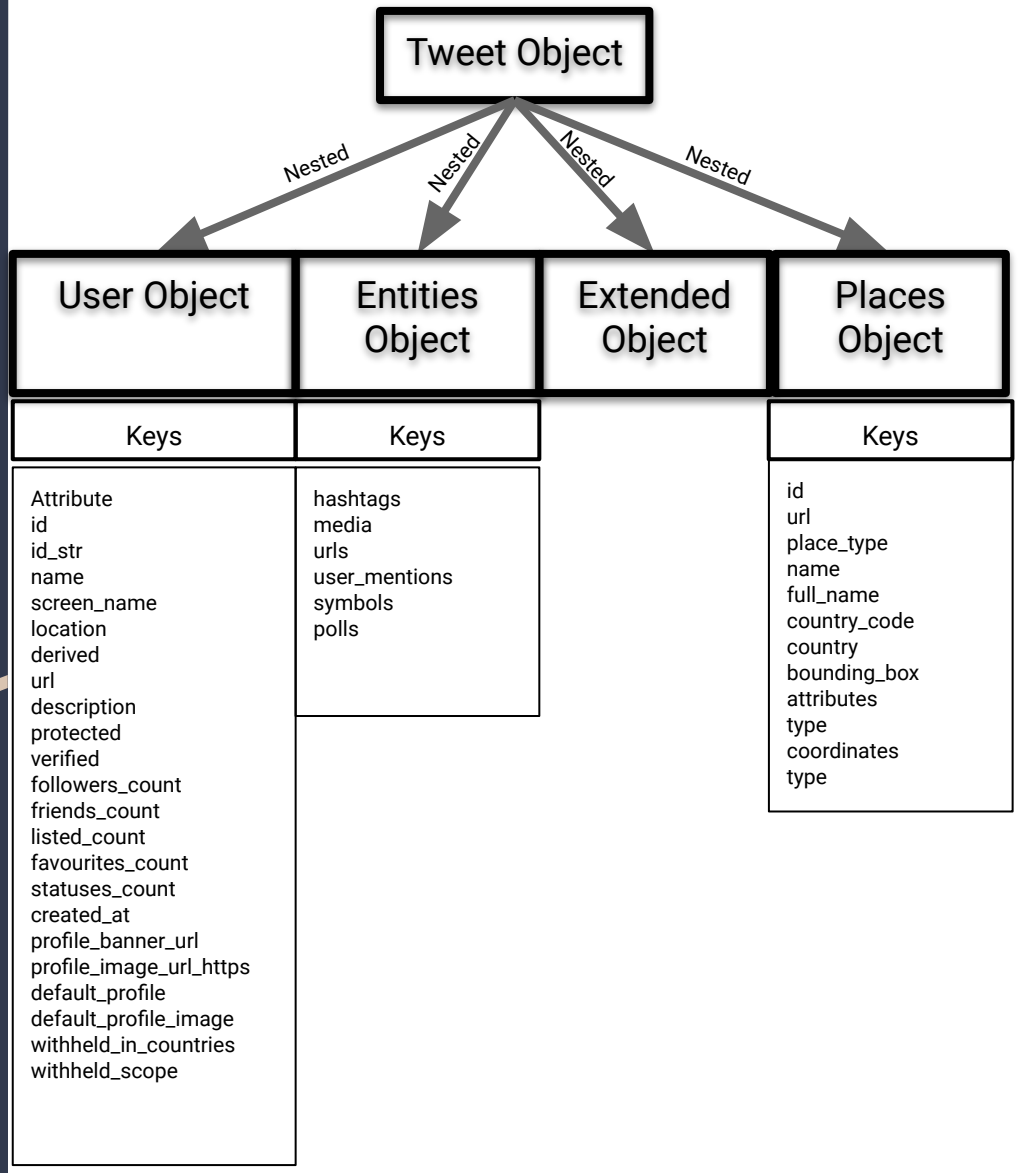
Data Acquisition

Data Acquisition Pseudo Code (Game Plan)

1. Applied for Twitter Developers Account
2. Received Secret Keys
3. Established oauth2 Token
4. Created Twitter Application using Python
5. Called Twitter Application Programming Interface (API)
6. Retrieved relevant Twitter Data

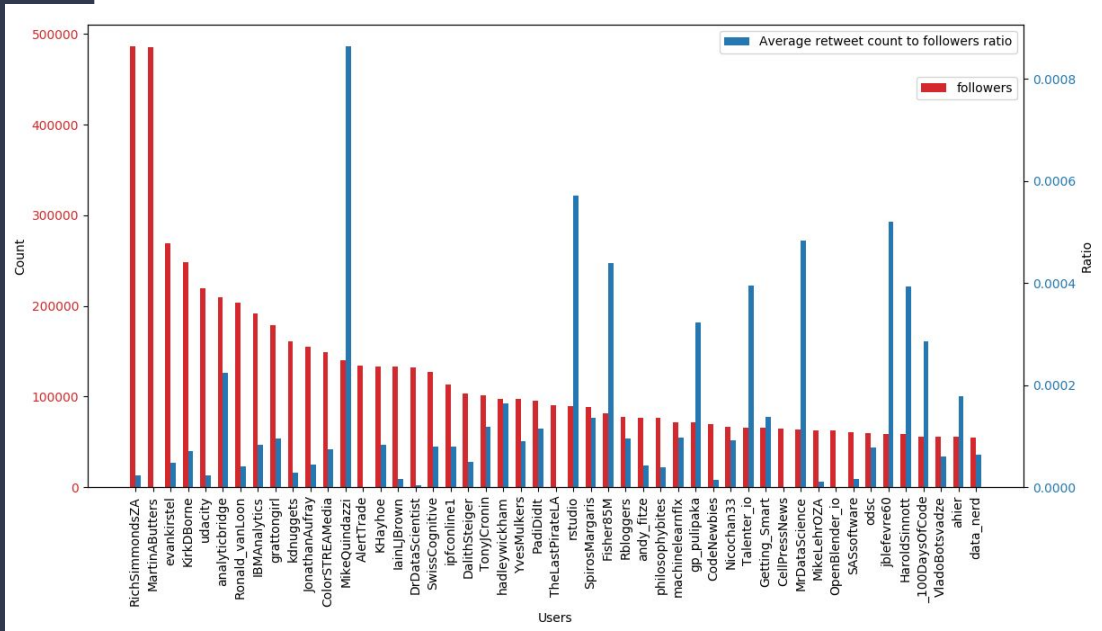
Twitter Data Structure

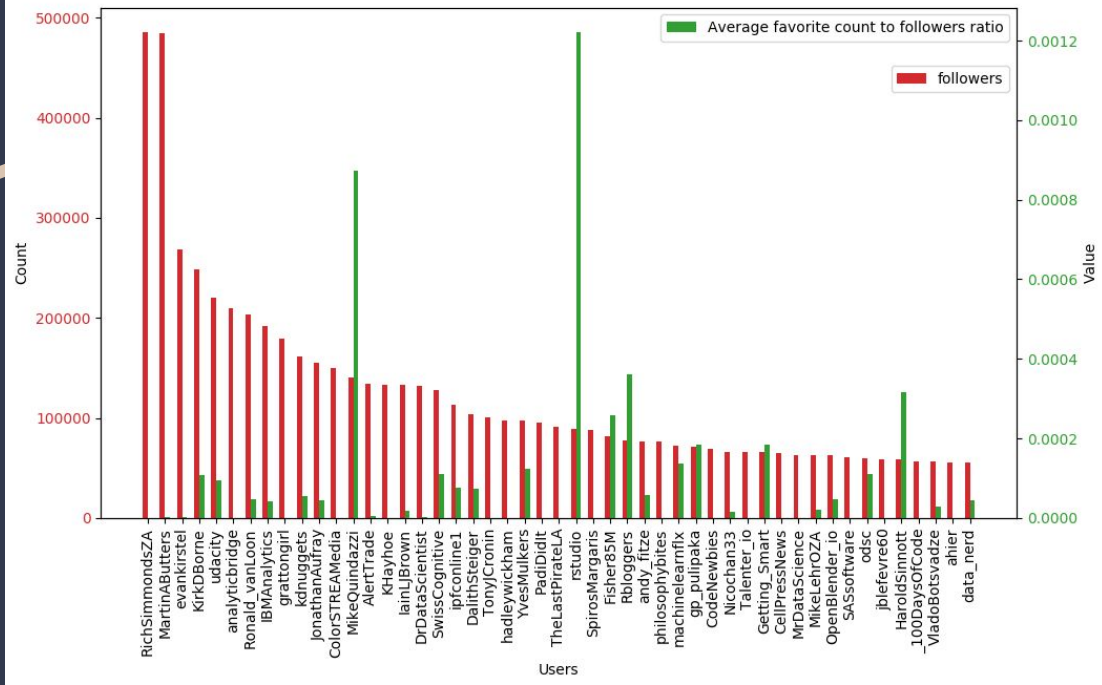
JavaScript Object Notation (JSON)



Influential Users

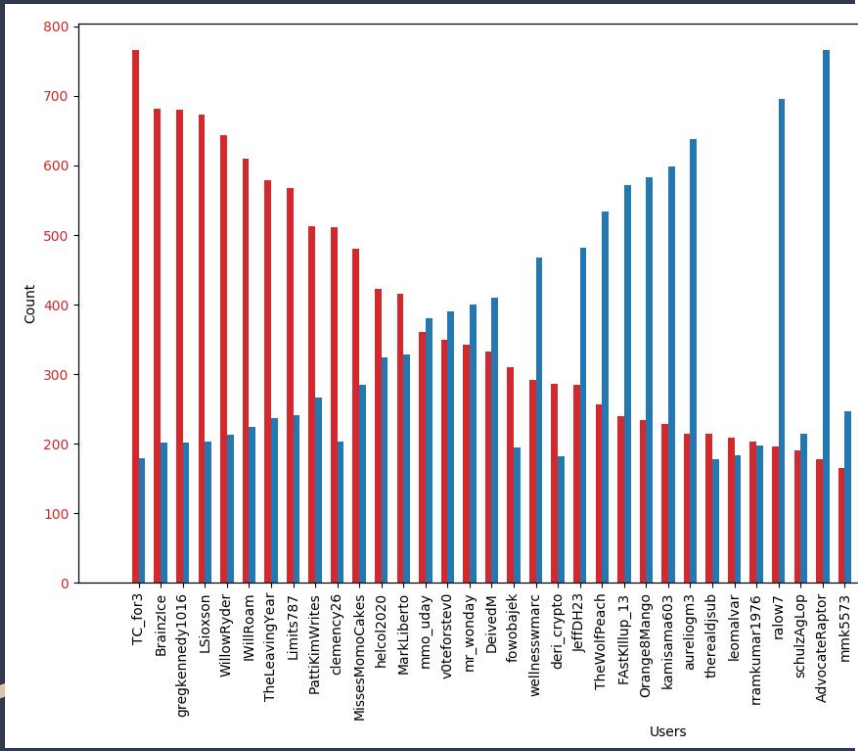
- Account for inactive followers.
- rf_ratio: Dividing the average number of retweets a user received by the number of followers the same user has.
- ff_ratio: Dividing the average number of favorites a user received by the number of followers the same user has.

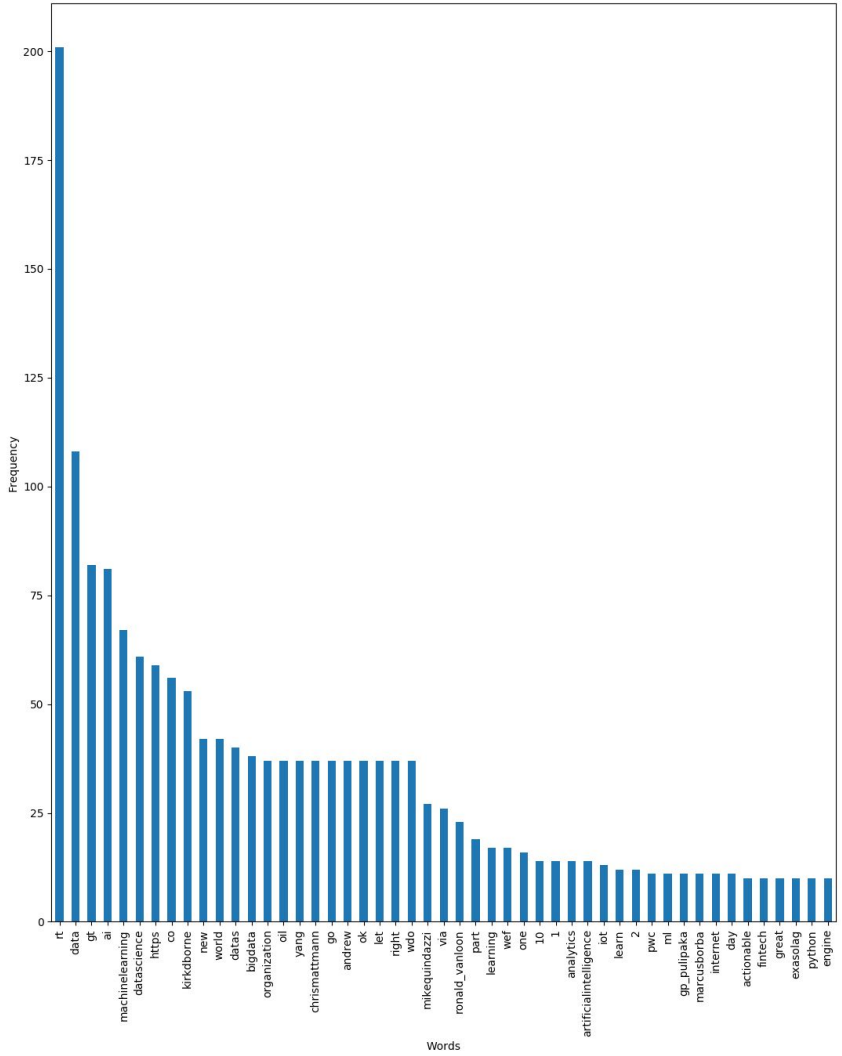




Influential Users

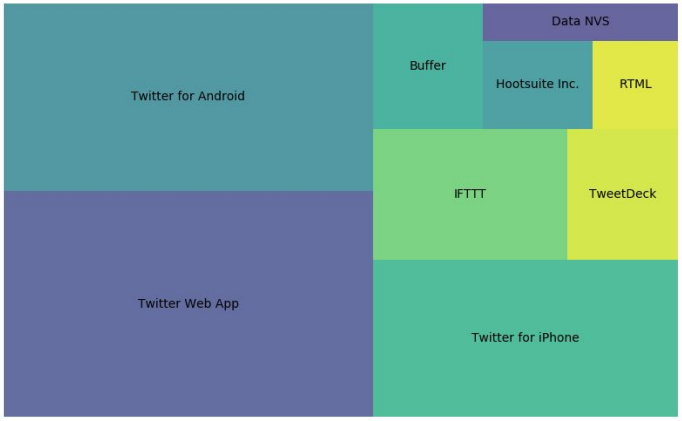
- Users with rf_ratio greater or equal to 1.
- What are these users tweeting about?

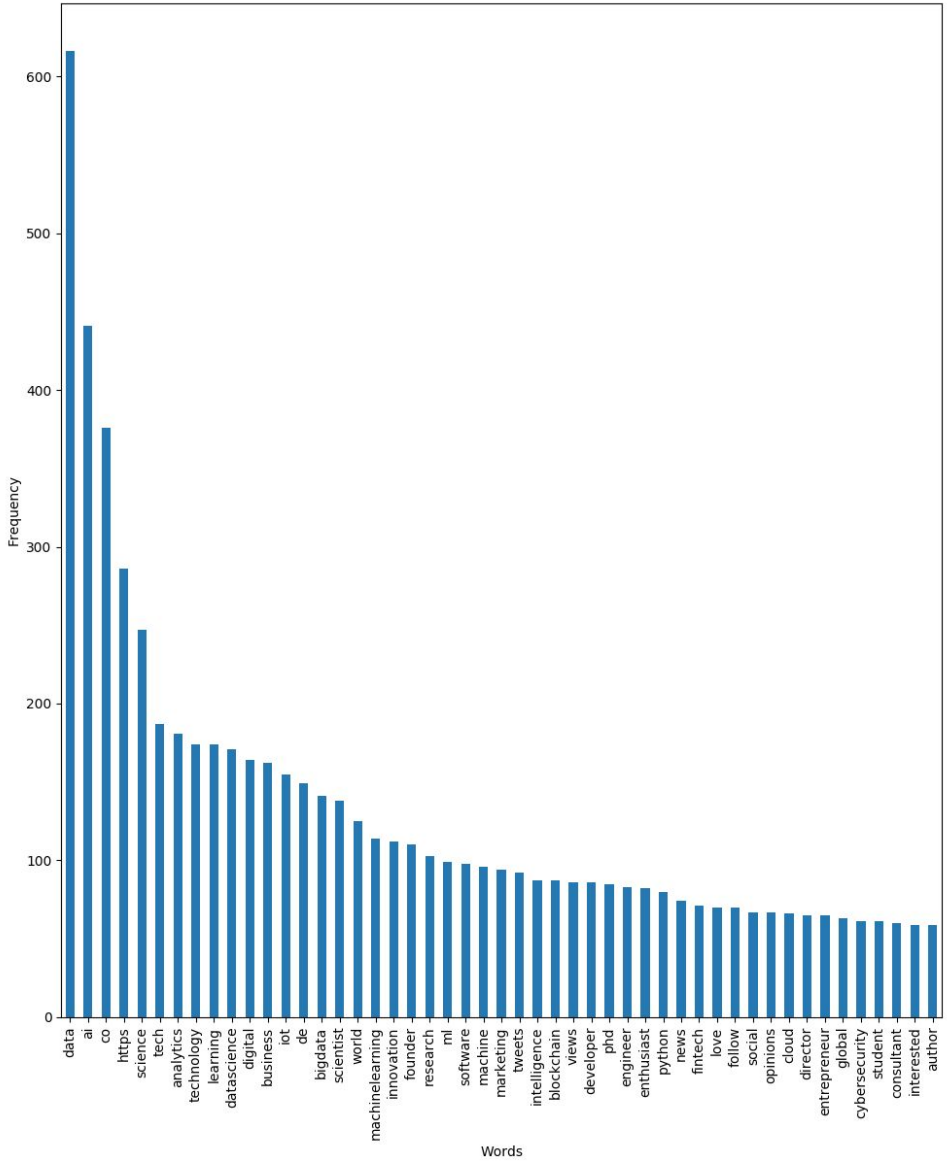




Influential Users

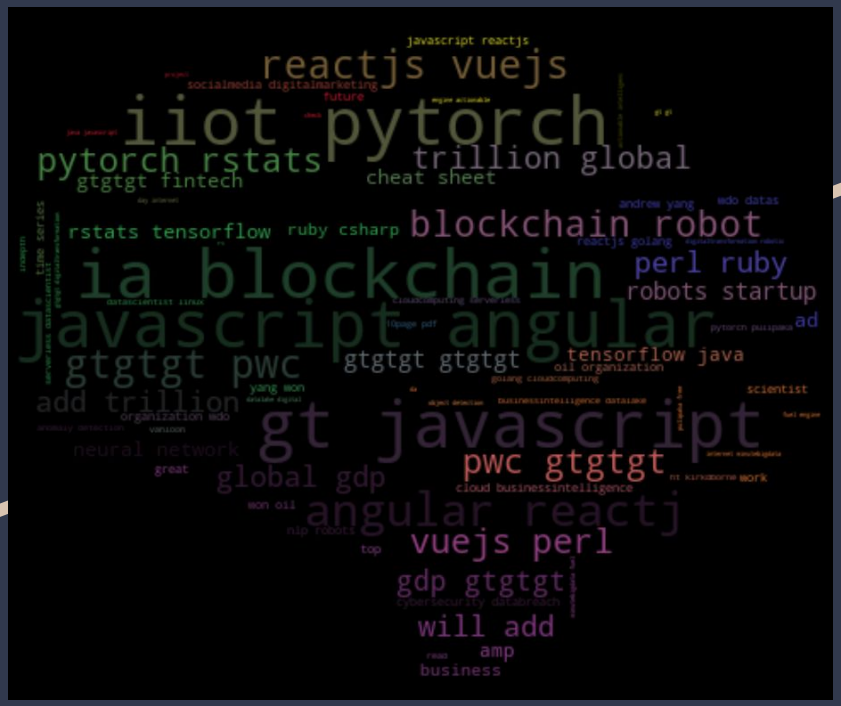
- How did these users describe themselves?
- What device did these users make the tweets from?





Influential Users

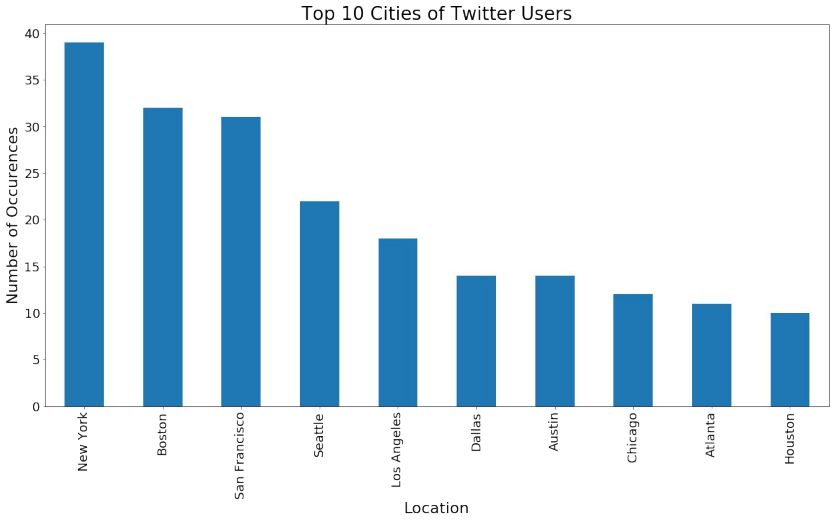
- Identified 3 main categories in the top followed users.
- Selected a few of these users to quarry new data set.
- What are these users talking about?

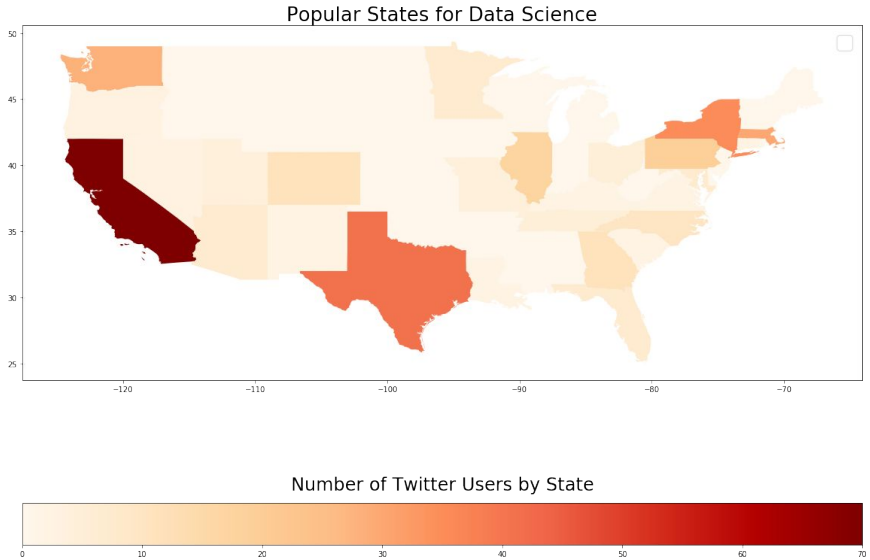


Influencer		Practitioner		Organization	
gt	732	https	999	https	249
https	271	co	999	co	249
co	271	datascience	443	data	69
mikequindazzi	125	bigdata	429	learning	33
via	123	ai	411	machine	25
ai	119	machinelearning	394	via	25
iot	114	python	269	wnxsdqrbfm	24
fisher85m	74	iot	247	ai	23
spirosmargaris	61	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

Location Analysis

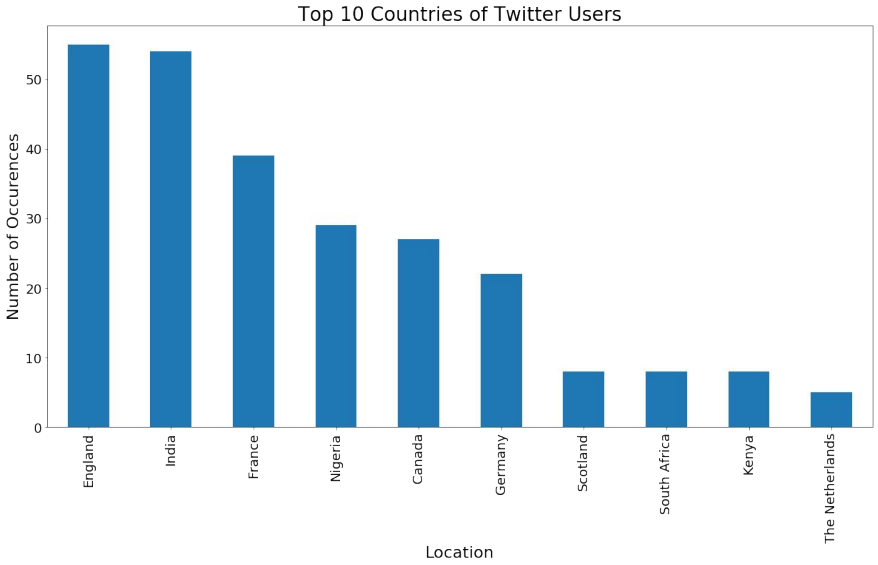
- Where are the twitter users who use the #DataScience?
- Data Cleaning Challenges
- U.S. Cities
  - New York, Boston, San Francisco
- U.S. States
  - California, Texas, New York

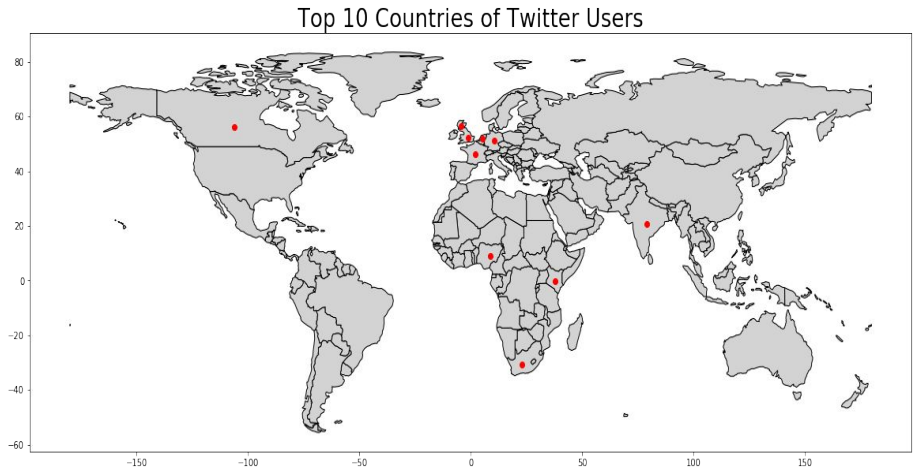




Location Analysis

- Data Cleaning Countries
  - England, London, UK, United Kingdom
- Europe
  - England, France, Germany
- Africa
  - Nigeria, Kenya, South Africa



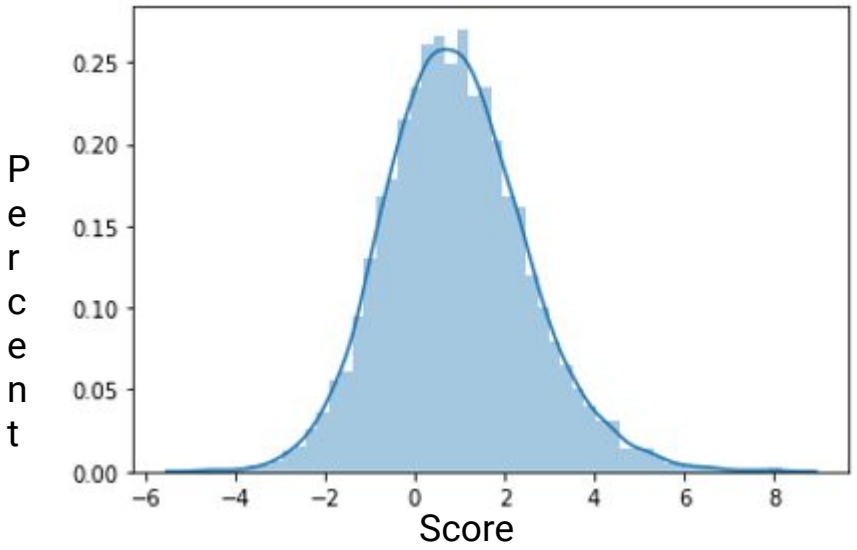


Sentiment Analysis

#DataScience

- Slightly Positive in Sentiment
- Methods
  - Count by Bag of Words
  - Semantic Orientation

Normal Distribution for Words using bag of words



count	mean	std	min	max
9000.0	0.927222	1.233912	-4	10

Semantic Orientation Scores

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

Thank You!
Any Questions?