2020-0701 IST652 Scripting for Data Analysis
HW # 2

Sathish Kumar Rajendiran

Prasad Kulkarni

2020-08-21

Table of Contents

Task Summary – Semi Structured Data Processing

Write a program that reads JSON formatted data (ex. Twitter Streaming data) from MongoDB collection. Analyze the data as lists of JSON structures and load them into Pandas dataframes for further analysis. Tasks includes (but not limited to),

1. Dataset Creation using Twitter API
    a. Setup developer account
    b. Authentication
    c. Define list of search keywords
2. Data Exploration
    a. Authentication into MongoDB & Objects creation
    b. Tweepy stream data into MongoDB
    c. Data Description
    d. Analyze JSON formatted collection
        i. MongoDB Compass
        ii. Python
    e. Pandas dataframes for processed data
3. Data Transformation
    a. Data type conversion
    b. Derive calendar items
    c. Remove redundant or unnecessary columns
    d. Remove special characters
    e. Re-arrange columns
    f. Write to file
4. Data Questions & Visualization
    a. Analyze twitter trends
        i. Top 10 users by followers
        ii. Ratio of percentage of Retweets vs Actual tweets
        iii. Top 10 retweets
        iv. Number of Tweets by day, hour of day, by month
    b. Word cloud
5. Summary
    a. What's the story(stories) in this data?
6. Conclusion
7. Contribution
    a. Task distribution

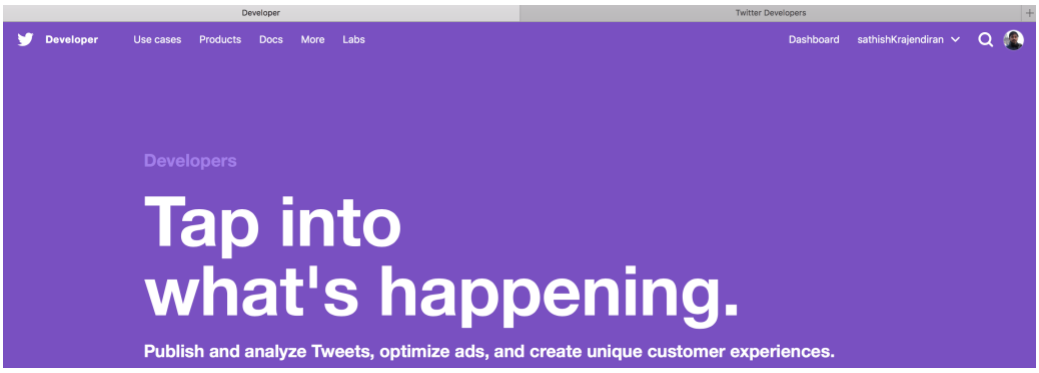
Data set creation using Twitter API

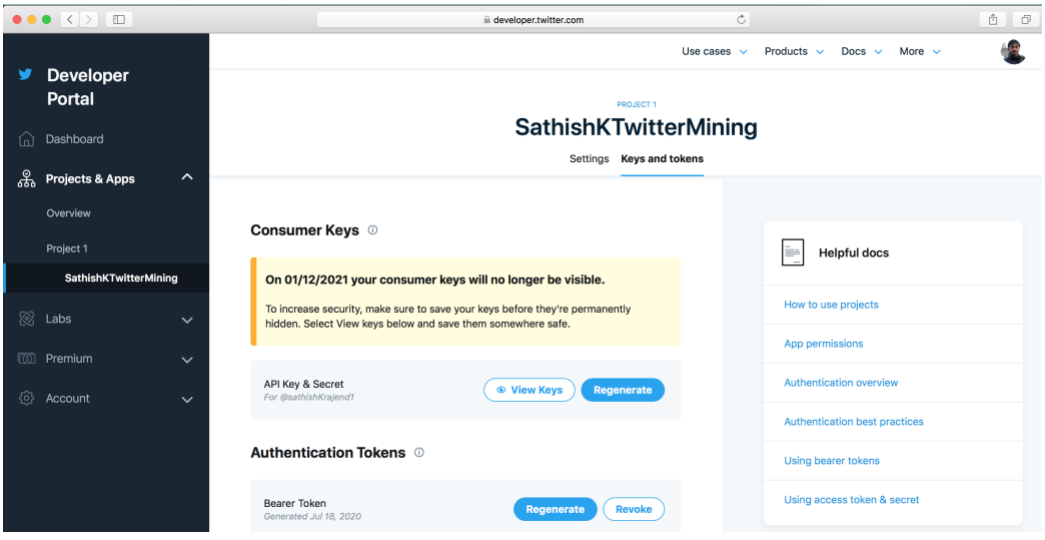
In order to perform tweets analysis, first step is to setup a developer account with Twitter. It involves the following steps,

a. Setup developer account
    1. Create developer account via https://developer.twitter.com/en
    2. Complete access request form by submitting a justification on why access being requested. Sample quote below,

        > "Master's in applied data science student of Prof. xxx xxx from Syracuse university is exploring semi structured data to analyze twitter feeds. Hence, access to twitter API is requested."

    3. Once developer account is activated, create an app and generate authentication token. This includes 4 key values,
        a. Consumer Key
        b. Consumer Secret
        c. Access Secret
        d. Access token





b. Authentication

Lets, authenticate through Jupyter Notebook using the Tweepy library from python and TwitterAPI keys. Credentials are read through an excel file (standard secret management) in the code.

```
Authentication OK - Youre now connected to the Twitter API.
```

c. Define list of search keywords

Let's define a list of keywords as "'#LA', '#LosAngeles', '#LAtraffic', '#LAFD', '#LASTREETCLOSURE'" to fetch tweets. Other parameters including language as "en"

Data Exploration

Next step is to stream twitter data into Mongo DB – a NoSQL database for semi structured (documents) data. Data exploration include the following steps,

a. Authentication into MongoDB & Objects creation

Initialize connection to Mongo DB local instance running on port number 27017. In addition, create "tweetsdb" and "tweets" collection object as well. This process will establish connection and create the databases objects if doesn't exist.

```
Authentication OK - Youre now connected to the MongoDB.

MongoDB database: Database(MongoClient(host=['localhost:27017'], document_class=dict, tz_aware=False, connect=True),
'tweetsdb')
MongoDB collection:Collection(Database(MongoClient(host=['localhost:27017'], document_class=dict, tz_aware=False, con
nect=True), 'tweetsdb'), 'tweets')
```

b. Tweepy stream data into MongoDB

Now that, Connections to Twitter API and Mongo DB are established; next step is to setup streaming listener and start collecting tweets into Mongo DB. This step includes, connecting to twitter feed and collects JSON formatted data and inserts into MongoDB collection "tweets"

```
Start Streaming...
Keywords:['#LA', '#LosAngeles', '#LAtraffic', '#LAFD', '#LASTREETCLOSURE']
Languages:['en']

Tweets follow...

 Now Playing Kirk Franklin - Love Theory  24/7 Christian Music and Live shows "Download the Anointed Radio App and c…
https://t.co/ANRVZ0lS9w

/opt/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:25: DeprecationWarning: insert is deprecated. Use in
sert_one or insert_many instead.

 #ExposingTheDemocrats #COUP against #TRUMP

Yea! 161.4k #Democrats have FLIPPED to #Republican in… https://t.co/n6xgqjAJe0

 Via @TravelChannel: "Top 10 #SouthernCalifornia Beaches." https://t.co/2eOIdfeQlz
```

c. Analyze JSON formatted collection

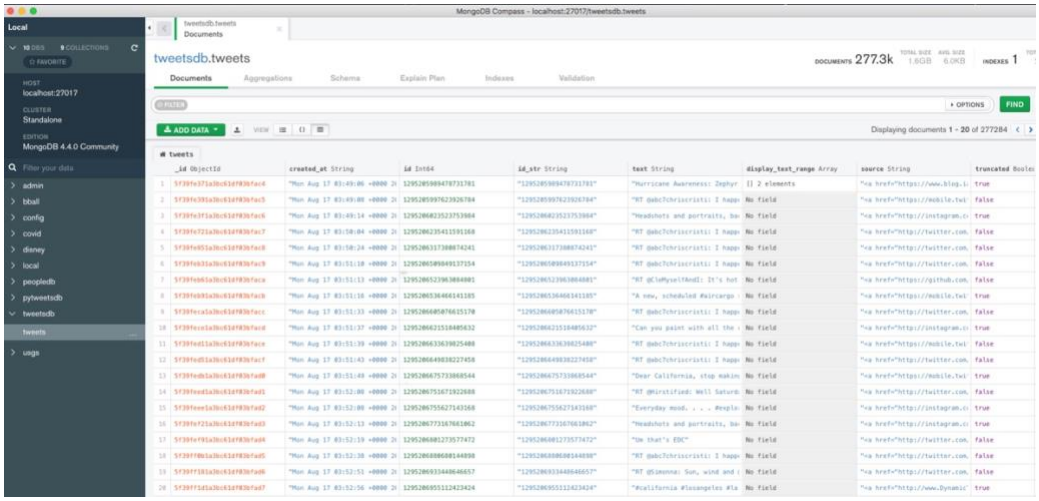
Let's review the tweets from MongoDB collection. It has over 277k tweets collected.

i. ipython Jupyter Notebook - First tweet from the collection as below,

{'_id': ObjectId('5f39fe371a3bc61df03bfac4'),
 'created_at': 'Mon Aug 17 03:49:06 +0000 2020',
 'id': 1295205989478731781,
 'id_str': '1295205989478731781',
 'text': 'Hurricane Awareness: Zephyr Insurance \n\nREAD MORE: https://t.co/SEhxK4bwyi\n\n#Accidents #Claims #DisasterMitigation… https://t.co/ew7ylyeHCz',
 'display_text_range': [0, 140],
 'source': 'IAMBLOG2TWITTER',
 'truncated': True,
 'in_reply_to_status_id': None,
 'in_reply_to_status_id_str': None,
 'in_reply_to_user_id': None,
 'in_reply_to_user_id_str': None,
 'in_reply_to_screen_name': None,
 'user': {'id': 226310002,
  'id_str': '226310002',
  'name': 'IAM Platform',
  'screen_name': 'IAM__Network',
  'location': 'Worldwide',
  'url': 'https://www.iammarketingmedia.com',
  'description': 'Curation | Tools | Tips | Services\n\nIAM Platform powers IAM Network:\n\nGO: http://bit.ly/2Ywsbg8\n\nBlog | Social | Podcast | Code Trove',
  'translator_type': 'none',
  'protected': False,
  'verified': False,
  'followers_count': 18016,
  'friends_count': 14938,
  'listed_count': 3290,
  'favourites_count': 65467,
  'statuses_count': 778665,
  'created_at': 'Mon Dec 13 21:24:29 +0000 2010',
  'utc_offset': None,
  'time_zone': None,
  'geo_enabled': False,
  'lang': None,
  'contributors_enabled': False,
  'is_translator': False,
  'profile_background_color': '94D487',
  'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme1/bg.png',
  'profile_background_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg.png',
  'profile_background_tile': False,
  'profile_link_color': '3366CC',
  'profile_sidebar_border_color': 'FFFFFF',
  'profile_sidebar_fill_color': 'DDEEF6',
  'profile_text_color': '333333',
  'profile_use_background_image': True,
  'profile_image_url': 'http://pbs.twimg.com/profile_images/701708113653669888/Nzm67hhC_normal.png',
  'profile_image_url_https': 'https://pbs.twimg.com/profile_images/701708113653669888/Nzm67hhC_normal.png',
  'profile_banner_url': 'https://pbs.twimg.com/profile_banners/226310002/1584072260',
  'default_profile': False,
  'default_profile_image': False,
  'following': None,
  'follow_request_sent': None,
  'notifications': None},
 'geo': None,
 'coordinates': None,
 'place': None,
 'contributors': None,
 'is_quote_status': False,
 'extended_tweet': {'full_text': 'Hurricane Awareness: Zephyr Insurance \n\nREAD MORE: https://t.co/SEhxK4bwyi\n\n#Accidents #Claims #DisasterMitigation #Insurance #InsuranceTechnology #InsurTech #Points #RiskMitigation #Technology~ https://t.co/kCrl2YxHfK',

```
 'display_text_range': [0, 194],
 'entities': {'hashtags': [{'text': 'Accidents', 'indices': [76, 86]},
   {'text': 'Claims', 'indices': [87, 94]},
   {'text': 'DisasterMitigation', 'indices': [95, 114]},
   {'text': 'Insurance', 'indices': [115, 125]},
   {'text': 'InsuranceTechnology', 'indices': [126, 146]},
   {'text': 'InsurTech', 'indices': [147, 157]},
   {'text': 'Points', 'indices': [158, 165]},
   {'text': 'RiskMitigation', 'indices': [166, 181]},
   {'text': 'Technology', 'indices': [182, 193]}],
  'urls': [{'url': 'https://t.co/SEhxK4bwyi',
    'expanded_url': 'https://blog.iammarketingmedia.com/hurricane-awareness-zephyr-insurance/?utm_campaign=twit
ter&utm_medium=twitter&utm_source=twitter',
    'display_url': 'blog.iammarketingmedia.com/hurricane-awar…',
    'indices': [51, 74]}],
  'user_mentions': [],
  'symbols': [],
  'media': [{'id': 1295205987209621505,
    'id_str': '1295205987209621505',
    'indices': [195, 218],
    'media_url': 'http://pbs.twimg.com/media/Efl--6qWsAEpwkT.jpg',
    'media_url_https': 'https://pbs.twimg.com/media/Efl--6qWsAEpwkT.jpg',
    'url': 'https://t.co/kCrl2YxHfK',
    'display_url': 'pic.twitter.com/kCrl2YxHfK',
    'expanded_url': 'https://twitter.com/IAM__Network/status/1295205989478731781/photo/1',
    'type': 'photo',
    'sizes': {'small': {'w': 448, 'h': 252, 'resize': 'fit'},
     'thumb': {'w': 150, 'h': 150, 'resize': 'crop'},
     'medium': {'w': 448, 'h': 252, 'resize': 'fit'},
     'large': {'w': 448, 'h': 252, 'resize': 'fit'}}}]},
 'extended_entities': {'media': [{'id': 1295205987209621505,
    'id_str': '1295205987209621505',
    'indices': [195, 218],
    'media_url': 'http://pbs.twimg.com/media/Efl--6qWsAEpwkT.jpg',
    'media_url_https': 'https://pbs.twimg.com/media/Efl--6qWsAEpwkT.jpg',
    'url': 'https://t.co/kCrl2YxHfK',
    'display_url': 'pic.twitter.com/kCrl2YxHfK',
    'expanded_url': 'https://twitter.com/IAM__Network/status/1295205989478731781/photo/1',
    'type': 'photo',
    'sizes': {'small': {'w': 448, 'h': 252, 'resize': 'fit'},
     'thumb': {'w': 150, 'h': 150, 'resize': 'crop'},
     'medium': {'w': 448, 'h': 252, 'resize': 'fit'},
     'large': {'w': 448, 'h': 252, 'resize': 'fit'}}}]},
 'quote_count': 0,
 'reply_count': 0,
 'retweet_count': 0,
 'favorite_count': 0,
 'entities': {'hashtags': [{'text': 'Accidents', 'indices': [76, 86]},
   {'text': 'Claims', 'indices': [87, 94]},
   {'text': 'DisasterMitigation', 'indices': [95, 114]}],
  'urls': [{'url': 'https://t.co/SEhxK4bwyi',
    'expanded_url': 'https://blog.iammarketingmedia.com/hurricane-awareness-zephyr-insurance/?utm_campaign=twit
ter&utm_medium=twitter&utm_source=twitter',
    'display_url': 'blog.iammarketingmedia.com/hurricane-awar…',
    'indices': [51, 74]},
   {'url': 'https://t.co/ew7ylyeHCz',
    'expanded_url': 'https://twitter.com/i/web/status/1295205989478731781',
    'display_url': 'twitter.com/i/web/status/1…',
    'indices': [116, 139]}],
  'user_mentions': [],
  'symbols': []},
 'favorited': False,
 'retweeted': False,
 'possibly_sensitive': False,
 'filter_level': 'low',
 'lang': 'en',
 'timestamp_ms': '1597636146312'}
```

ii. MongoDB Compass – view the collection from MongoDB UI Interface (Compass). This interface provides better view of the JSON formatted data in both tabular and JSON format.




```
_id: ObjectId("5f39fe371a3bc61df03bfac4")
created_at: "Mon Aug 17 03:49:06 +0000 2020"
id: 1295205989478731781
id_str: "1295205989478731781"
    "Hurricane Awareness: Zephyr Insurance
text:
    READ MORE: https://t.co/SEhxK4..."
display_text_range: Array
source: "<a href="https://www.blog.iammarketingmedia.com" rel="nofollow">IAMBLO..."
truncated: true
in_reply_to_status_id: null
in_reply_to_status_id_…: null
in_reply_to_user_id: null
in_reply_to_user_id_str: null
in_reply_to_screen_name: null
user: Object
  id: 226310002
  id_str: "226310002"
  name: "IAM Platform"
  screen_name: "IAM__Network"
  location: "Worldwide"
  url: "https://www.iammarketingmedia.com"
            "Curation | Tools | Tips | Services

  description: IAM Platform powers IAM Network:

            ..."
  translator_type: "none"
  protected: false
  verified: false
  followers_count: 18016
  friends_count: 14938
  listed_count: 3290
  favourites_count: 65467
  statuses_count: 778665
  created_at: "Mon Dec 13 21:24:29 +0000 2010"
  utc_offset: null
  time_zone: null
  geo_enabled: false
  lang: null
  contributors_enabled: false
  is_translator: false
  profile_background_col…: "94D487"
  profile_background_ima…: "http://abs.twimg.com/images/themes/theme1/bg.png"
  profile_background_ima…: "https://abs.twimg.com/images/themes/theme1/bg.png"
  profile_background_tile: false
  profile_link_color: "3366CC"
  profile_sidebar_border…: "FFFFFF"
  profile_sidebar_fill_c…: "DDEEF6"
  profile_text_color: "333333"
  profile_use_background…: true
  profile_image_url: "http://pbs.twimg.com/profile_images/701708113653669888/Nzm67hhC_normal..."
                     "https://pbs.twimg.com/profile_images/701708113653669888/Nzm67hhC_norm
  profile_banner_url: "https://pbs.twimg.com/profile_banners/226310002/1584072260"
  default_profile: false
  default_profile_image: false
  following: null
  follow_request_sent: null
  notifications: null
geo: null
coordinates: null
place: null
contributors: null
is_quote_status: false
› extended_tweet: Object
quote_count: 0
reply_count: 0
retweet_count: 0
favorite_count: 0
› entities: Object
favorited: false
retweeted: false
possibly_sensitive: false
filter_level: "low"
lang: "en"
timestamp_ms: "1597636146312"
```

d. Data Description: Find the data description of the fields that are in scope here

Column	Description	Data Type	Range of Values
Id	Unique Identifier for this tweet	Integer	Ex. “id:105011842332”
Name	Username	String	Ex. “David”
Text	Actual UTF-8 Twitter text	String	Ex. “David Reese: #WaterWorld #Federal ism SoCalGas sues California over climate change policy - Los Angeles Times”
Coordinates	Geographic location of this tweet as reported	Coordinates	Ex. "coordinates": { "coordinates":[-75.14310264,40.05701649], "type":"Point"}
Retweet_count	Number of times this tweet has been retweeted	Integer	Ex."retweet_count":160
Source	Where the tweet is posted from	String	Ex. <a href >url reference. Instagram
Lang	Language Identifier	String	Ex. "lang": "en"
Favorite_count	Approximately how many times this tweet has been liked	Integer	Ex. "favorite_count":295
Followers_count	Approximately how many followers that the user has	Integer	Ex. "followers_count":500
timestamp_ms	UTC time when this tweet was created	String	Time values in Unix “1597636146312”

e. Pandas dataframes for processed data

Now, let’s collect few selected fields into a list and load it into pandas dataframes for further analysis. Fields in scope are id, id_str, username, source, followers_count, retweets_count, coordinates, place, full_name, text, profile_background_color and possibly_sensitive are collected through tw_list [] as below. Later, data from tw_list [] is loaded into “tweetsDF” pandas dataframe.

tweetsDF.info displays more information about the dataframe including Number of rows as in index range, missing values, column name, datatype and total memory allocated.

```
1 #Data Processing on tweetsDF
2 tweetsDF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 277273 entries, 0 to 277272
Data columns (total 10 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   id          277273 non-null  int64
 1   id_str      277273 non-null  object
 2   user        277273 non-null  object
 3   source      277273 non-null  object
 4   followers   277273 non-null  int64
 5   retweets    277273 non-null  int64
 6   coords      1189 non-null    object
 7   bg_color    277273 non-null  object
 8   datets      277273 non-null  object
 9   text        277273 non-null  object
dtypes: int64(3), object(7)
memory usage: 21.2+ MB
```

```
1 #Analyze Dataframe - shape
2 tweetsDF.shape
```

```
(277284, 10)
```

tweetsDF.head() displays top 5 rows from the dataframe as in rows and columns.

```
# Change the column names
# tweetsDF.columns = ['id','UserName','Source','DateTime','Text']
tweetsDF.head()
```

user	source	followers	retweets	coords	bg_color	datets	text
Platform	IAMBLOG2TWITTER	18016	0	None	94D487	2020-08-16 20:49:06	Hurricane Awareness: Zephyr Insurance \n\nREAD MORE: https://t.co/SEhxK4bwyi\n\n#Accidents #Claims #DisasterMitigation #Insurance #InsuranceTechnology #InsurTech #Points #RiskMitigation #Technology~ https://t.co/kCrl2YxHfK
son(Ally) reeman, Day)	Twitter Web App	298	0	None	C0DEED	2020-08-16 20:49:08	RT @abc7chriscristi: I happen to think LAX is uniquely photogenic at night... for an airport. #air7hd @abc7 #abc7eyewitness #lax #airport #…
DeAndre n Media	Instagram	216	0	{'type': 'Point', 'coordinates': [-118.2445, 34.0564]}	9AE4E8	2020-08-16 20:49:14	Headshots and portraits, back open for business. DM for bookings. \n\n ______________________ •\n•\n•\n\n#sony #modellife #fashionblogger #sonyalpha #fashion #losangeles #travelingphotographer #california #newyork… https://t.co/CmzulO1JgK
Valverde	Twitter for iPhone	743	0	None	000000	2020-08-16 20:50:04	RT @abc7chriscristi: I happen to think LAX is uniquely photogenic at night... for an airport. #air7hd @abc7 #abc7eyewitness #lax #airport #…
Alfredo	Twitter for iPad	354	0	None	1A1B1F	2020-08-16 20:50:24	RT @abc7chriscristi: I happen to think LAX is uniquely photogenic at night... for an airport. #air7hd @abc7 #abc7eyewitness #lax #airport #…

Data Transformation

Preprocessing complete. Next step is to prepare the data for further analysis. It involves, cleaning the data, data type conversion, removal of unwanted columns, removal of special characters, creation of new columns based on calendar date time, re-arranging the column sequence and exporting it to csv file.

a. Data type conversion & Derive calendar items

Let's convert "datets" field to datetime datatype from defaulted "object" datatype and create additional columns from this field such as Date, year, Month, Monthday, weekday, monthname, hour, minute and seconds for tweets analysis at different calendar date/time schedules.

bg_color	datets	text	date	year	month	monthday	weekday	dayname	monthname	hour	minute	secs
94D487	2020-08-16 20:49:06	Hurricane Awareness: Zephyr Insurance \n\nREAD MORE: https://t.co/SEhxK4bwyi\n\n#Accidents #Claims #DisasterMitigation #Insurance #InsuranceTechnology #InsurTech #Points #RiskMitigation #Technology~ https://t.co/kCrl2YxHfK	2020-08-16	2020	8	16	6	Sunday	August	20	49	6
C0DEED	2020-08-16 20:49:08	RT @abc7chriscristi: I happen to think LAX is uniquely photogenic at night... for an airport. #air7hd @abc7 #abc7eyewitness #lax #airport #…	2020-08-16	2020	8	16	6	Sunday	August	20	49	8
9AE4E8	2020-08-16 20:49:14	Headshots and portraits, back open for business. DM for bookings. \n\n •\n•\n•\n\n#sony #modellife #fashionblogger #sonyalpha #fashion #losangeles #travelingphotographer #california #newyork… https://t.co/CmzulO1JgK	2020-08-16	2020	8	16	6	Sunday	August	20	49	14
000000	2020-08-16 20:50:04	RT @abc7chriscristi: I happen to think LAX is uniquely photogenic at night... for an airport. #air7hd @abc7 #abc7eyewitness #lax #airport #…	2020-08-16	2020	8	16	6	Sunday	August	20	50	4
1A1B1F	2020-08-16 20:50:24	RT @abc7chriscristi: I happen to think LAX is uniquely photogenic at night... for an airport. #air7hd @abc7 #abc7eyewitness #lax #airport #…	2020-08-16	2020	8	16	6	Sunday	August	20	50	24

b. Remove special characters

Next step is to cleanup "source" column by removing all the special characters and retain only the platform detail into a new column.

	text	date	year	month	monthday	weekday	dayname	monthname	hour	minute	secs	platform
	Hurricane Awareness: Zephyr Insurance \n\nREAD MORE: https://t.co/SEhxK4bwyl\n\n#Accidents #Claims #DisasterMitigation #Insurance #InsuranceTechnology #InsurTech #Points #RiskMitigation #Technology~ https://t.co/kCrl2YxHfK	2020-08-16	2020	8	16	6	Sunday	August	20	49	6	IAMBLOG2TWITTER
	RT @abc7chriscristi: I happen to think LAX is uniquely photogenic at night... for an airport. #air7hd @abc7 #abc7eyewitness #lax #airport #…	2020-08-16	2020	8	16	6	Sunday	August	20	49	8	Twitter Web App
	Headshots and portraits, back open for business. DM for bookings. \n\n ———————————————————— \n•\n•\n•\n\n#sony #modellife #fashionblogger #sonyalpha #fashion #losangeles #travelingphotographer #california #newyork… https://t.co/CmzulO1JgK	2020-08-16	2020	8	16	6	Sunday	August	20	49	14	Instagram
	RT @abc7chriscristi: I happen to think LAX is uniquely photogenic at night... for an airport. #air7hd @abc7 #abc7eyewitness #lax #airport #…	2020-08-16	2020	8	16	6	Sunday	August	20	50	4	Twitter for iPhone
	RT @abc7chriscristi: I happen to think LAX is uniquely photogenic at night... for an airport. #air7hd @abc7 #abc7eyewitness #lax #airport #…	2020-08-16	2020	8	16	6	Sunday	August	20	50	24	Twitter for iPad

c. Remove redundant or unnecessary columns

Columns “id_str”, “source”, “retweets” are removed as no longer needed as they seem redundant.

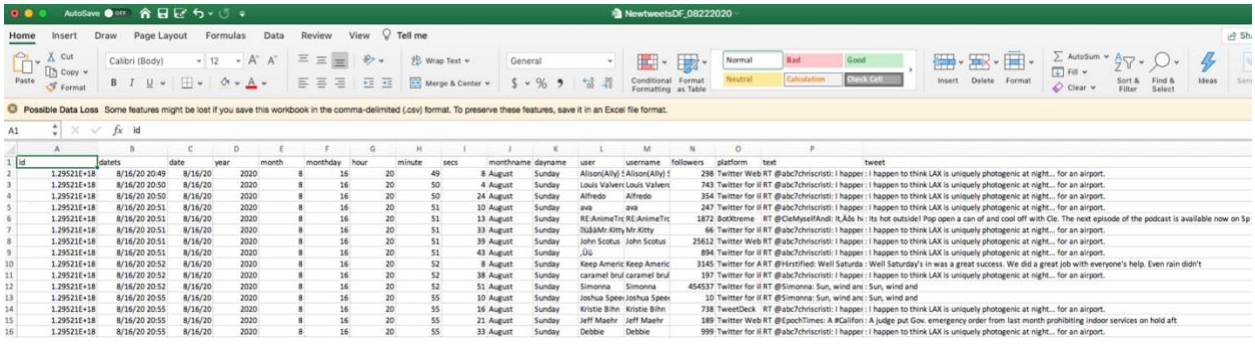
d. Preprocessing API is used to clean Text and userName fields with special characters for tokenization.

e. Re-arrange the columns

Let’s rearrange the columns for ease of analysis.

	id	datets	date	year	month	monthday	hour	minute	secs	monthname	dayname	user	username	followers
0	1295205989478731781	2020-08-16 20:49:06	2020-08-16	2020	8	16	20	49	6	August	Sunday	IAM Platform	IAM Platform	18016
1	1295205997623926784	2020-08-16 20:49:08	2020-08-16	2020	8	16	20	49	8	August	Sunday	Alison(Ally) Stapf(Freeman, Day)	Alison(Ally) Stapf(Freeman, Day)	298
2	1295206023523753984	2020-08-16 20:49:14	2020-08-16	2020	8	16	20	49	14	August	Sunday	DeAndre Jackson Media	DeAndre Jackson Media	216
3	1295206235411591168	2020-08-16 20:50:04	2020-08-16	2020	8	16	20	50	4	August	Sunday	Louis Valverde	Louis Valverde	743
4	1295206317380874241	2020-08-16 20:50:24	2020-08-16	2020	8	16	20	50	24	August	Sunday	Alfredo	Alfredo	354

f. Write to file – export the file into CSV for validation



Data Questions & Visualizations

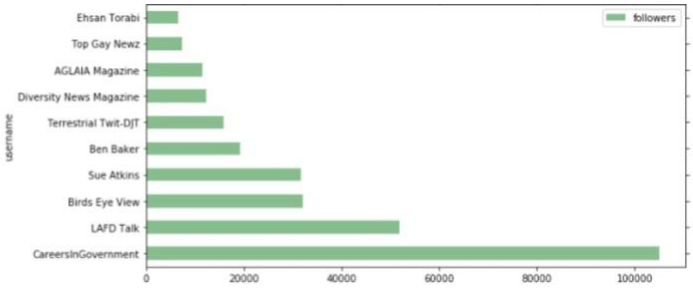
Let's explore twitter feeds further; For the analysis only focus on text containing "las" string.

```
# ********************************************************************************
#       Create a Pandas data frame with filter words keywords to analyze  LA traffic
# ********************************************************************************

# create a data frame containing text "las" from tweetsDF
la_tweets = NewtweetsDF[NewtweetsDF['text'].str.contains('los angeles|latraffic|california|LAFD|LAPD')]

# la_tweets.shape
# la_tweets.head()
```

a. Analyze twitter trends

i. Top 10 users by followers – from the above dataframe lets find out top 10 users by their number of followers.

```
Top 10 users by followers:

CareersInGovernment :  105,108
LAFD Talk :  51,768
Birds Eye View :  32,004
Sue Atkins :  31,767
Ben Baker :  19,160
Terrestrial Twit-DJT :  15,833
Diversity News Magazine :  12,288
AGLAIA Magazine :  11,530
Top Gay Newz :  7,424
Ehsan Torabi :  6,535
```



ii. Find min, max, average number of followers

```
Maximum Number of followers:  105108
Avergage Number of followers:  3029.0
Minimum Number of followers:  0
```

iii. Ratio of percentage of Retweets vs Actual tweets

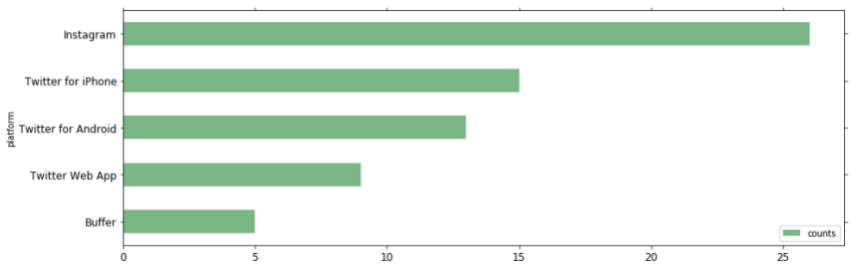
From the revised dataframe, lets find out ratio of retweets vs direct tweets

```
Percentage of retweets 45.06%
Percentage of actual tweets 54.94%
```

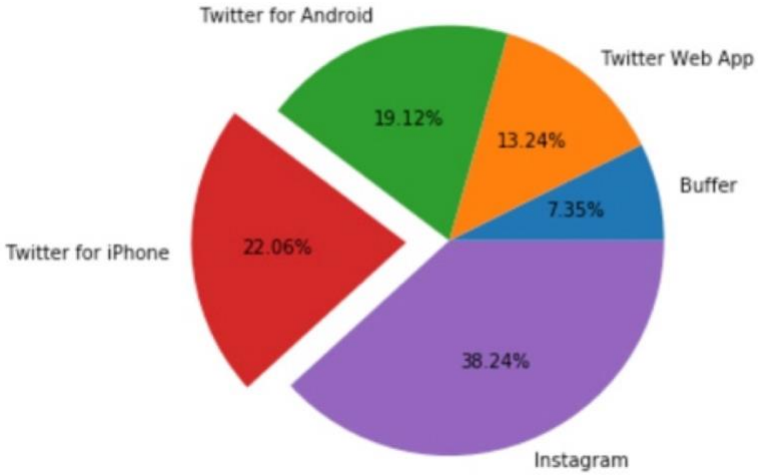
iv. Top 10 tweets by platform

```
Top 5 Platform by users:
```

	platform	counts
1	Instagram	26
2	Twitter for iPhone	15
3	Twitter for Android	13
4	Twitter Web App	9
5	Buffer	5

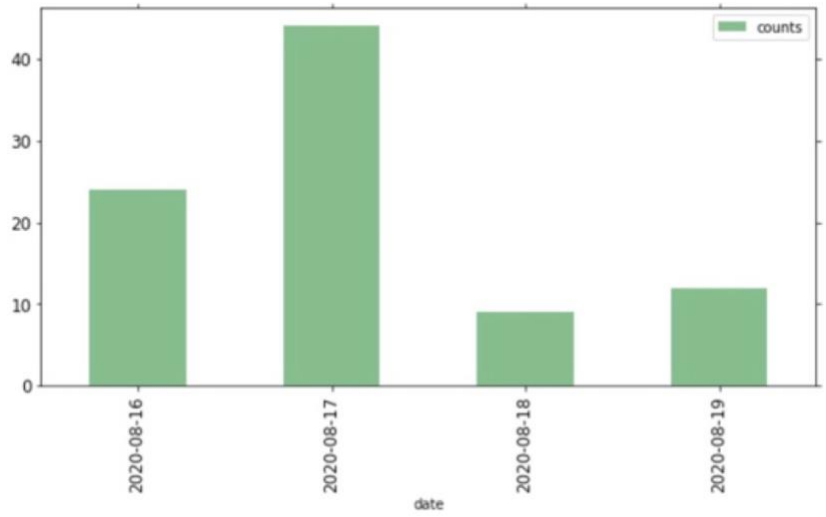


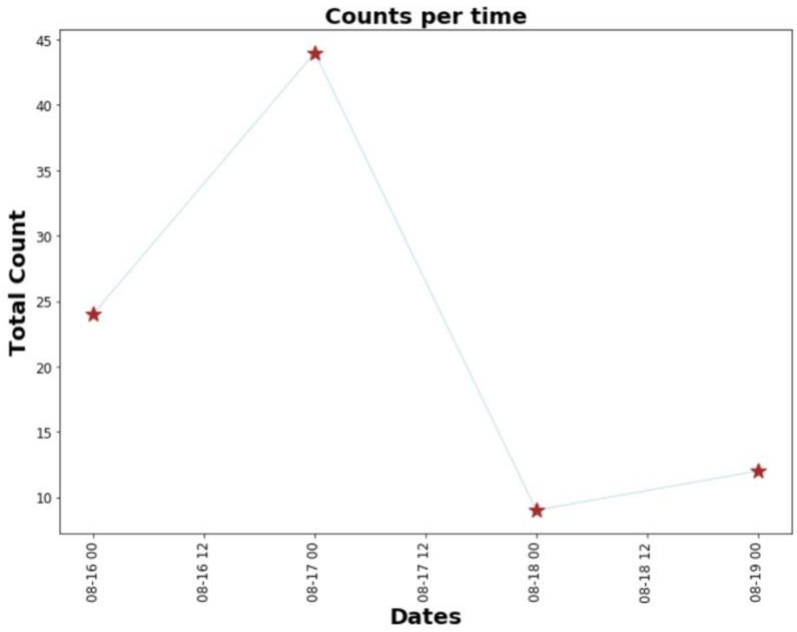
Top 5 Platform by users



v. Number of Tweets by day, hour of day, by month

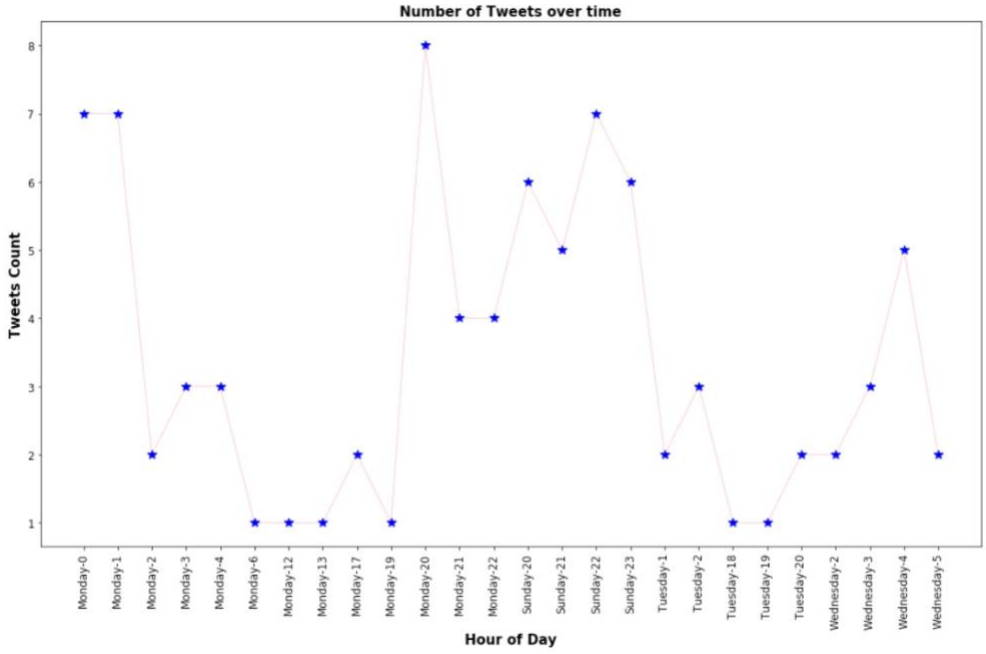
Tweets by calender date:





Tweets by week day of the Month:

	monthname	dayname	counts
1	August	Monday	44
2	August	Sunday	24
3	August	Tuesday	9
4	August	Wednesday	12



Tweets by hour of the day:

	dayname	hour	counts
1	Monday	0	7
2	Monday	1	7
3	Monday	2	2
4	Monday	3	3
5	Monday	4	3
6	Monday	6	1
7	Monday	12	1
8	Monday	13	1
9	Monday	17	2
10	Monday	19	1
11	Monday	20	8
12	Monday	21	4
13	Monday	22	4
14	Sunday	20	6
15	Sunday	21	5
16	Sunday	22	7
17	Sunday	23	6
18	Tuesday	1	2
19	Tuesday	2	3
20	Tuesday	18	1
21	Tuesday	19	1
22	Tuesday	20	2
23	Wednesday	2	2
24	Wednesday	3	3
25	Wednesday	4	5
26	Wednesday	5	2

b. Word cloud – Lastly, a Wordcloud on twitter feeds to conclude the trending words on the key words including #latraffic, #lapd, #lafd, #losangeles city – designed in a butterfly shape. It also excluded few common stop words like “https”,” rt”,” hi”,” promo”,” thank” etc.

Summary

After extracting data via Twitter API, following observations were made.

- Data is in JSON format.
- Tweepy streaming API pulled more than 270k tweets
- MongoDB datastore used to collect all the tweets
- Pandas dataframes used for data transformation, processing and answering data questions with simple plots (bar, line graph and pie chart)
- Simple Word cloud representation made in butterfly shape

- Analysis including
    - tweets filtered on keywords such as #losangeles, #latraffic, #lapd, #lafd etc.
    - 45% of the tweets were re-tweets
    - 38% of tweets were posted through Instagram followed by iPhone with 22%
    - Monday 6 AM showed the least number of tweets where as 8PM on Monday had the highest tweets traffic. Monday remained a busy day with more tweets by day, followed by Sunday.
    - CareersInGovernment user remained top with number of active followers; followed by LAFD talk
    - Average followers count is ~3000
    - Wordcloud highlighted some of the top words as "outpouring", "heavy hearts", "support", "jose perez" , "fire station" etc. – highlighting the outpouring support from users on twitter in support of "Jose Perez- a fire fighter who had lost his life due to COVID 19" https://abc30.com/coronavirus-deaths-firefighter-death-covid-19/6338926/

Conclusion

Overall, an amazing exercise to explore the power of twitter streaming api, how to use MongoDB to store and view JSON- Semi structured data collections effortlessly, Power of Pandas dataframe to perform data cleaning/transformation, analysis tasks with visualization options including plots/graphs/charts to answer any meaningful data questions.
Lastly, with over 1.3 billion users – twitter feeds can be considered as gold mine of data that can be used effectively to get insights of the general public on any topic that is trending. Based on the public's reaction to topics/products/promotional events – a success of marketing campaigns can be determined and even improvised based on target audience pulse.

Contributions

- Project Topic: Sathish Kumar Rajendiran
- Twitter Streaming: Sathish Kumar Rajendiran
- MongoDB & JSON Analysis: Prasad Kulkarni, Sathish Kumar Rajendiran
- Data Cleansing and Formatting: Prasad Kulkarni, Sathish Kumar Rajendiran
- Twitter Trend Analysis: Prasad Kulkarni
- Wordcloud: Sathish Kumar Rajendiran
- Summary & Conclusion: Prasad Kulkarni, Sathish Kumar Rajendiran