Twitter Sentiment Analyzer

By: Kahlil Wehmeyer, Richard Cruz, Luke Rhon, Diego De La Torre, and Jackie Gauthier

Project Description

- Build a python program that will take input about any subject and provide a twitter statistics on it.
- Program features:
  - Search Twitter for tweets that are related to that topic.
  - Perform sentiment analysis and some automatic exploratory data analysis.

Problem Description

- Twitter is practically a library of real world sentiment to a wide range of topics.
- For companies and individuals such as journalist and industry researchers it may be useful to get a wisdom of the crowd view of the topics they are researching.
  - Example: Tech companies such as Facebook, Google, and Amazon might want to see whether people have a positive, neutral, or negative opinion of their company

Searching Tweets

- Created a function that inputs a search query
- Return a mapped dataframe from the API JSON

```python
def search(query):

    tweets = api.search(q=query)

    DataSet = pd.DataFrame()

    DataSet['tweetID'] = [tweet.id for tweet in tweets]
    DataSet['tweetText'] = [tweet.text for tweet in tweets]
    DataSet['tweetRetweetCt'] = [tweet.retweet_count for tweet
    in tweets]
    DataSet['tweetFavoriteCt'] = [tweet.favorite_count for tweet
    in tweets]
    DataSet['tweetSource'] = [tweet.source for tweet in tweets]
    DataSet['tweetCreated'] = [tweet.created_at for tweet in tweets]


    DataSet['userID'] = [tweet.user.id for tweet in tweets]
    DataSet['userScreen'] = [tweet.user.screen_name for tweet
    in tweets]
    DataSet['userName'] = [tweet.user.name for tweet in tweets]
    DataSet['userCreateDt'] = [tweet.user.created_at for tweet
    in tweets]
    DataSet['userDesc'] = [tweet.user.description for tweet in tweets]
    DataSet['userFollowerCt'] = [tweet.user.followers_count for tweet
    in tweets]
    DataSet['userFriendsCt'] = [tweet.user.friends_count for tweet
    in tweets]
    DataSet['userLocation'] = [tweet.user.location for tweet in tweets]
    DataSet['userTimezone'] = [tweet.user.time_zone for tweet
    in tweets]

    return DataSet
```

Searching Tweets Example

```
1   Florida_Poly_Tweets = search("Florida Poly")
```

```
1   Florida_Poly_Tweets
```

	tweetID	tweetText	tweetRetweetCt	tweetFavoriteCt	tweetSource	tweetCreated	userID	userScreen	userName
0	1106009631640481793	The Poly Prep Baseball Team will participate i...	0	4	Twitter for iPhone	2019-03-14 01:50:19	903309242551209988	PolyBlueDevils	Poly Blu Devil
1	1106006717857447936	RT @FLPolyU: More research opportunities are o...	1	0	Twitter for Android	2019-03-14 01:38:44	897831175846334464	CCCHillsborough	Debbie Wanne
2	1105987765160894464	Florida Poly unveils plans for state-of-the-ar...	0	0	IFTTT	2019-03-14 00:23:26	16233017	theledger	The Ledge
3	1105951677071319040	More research opportunities are on the horizon...	1	10	Zoho Social	2019-03-13 22:00:02	1321139442	FLPolyU	Florida Pol
4	1105887731039666177	The Phoenix🔥🔥🔥 tokyodachi of Florida 🌴 Poly.. ...	0	0	Instagram	2019-03-13 17:45:56	1029951734125219840	tokyodachi	Tokyodach
5	1105873389712015365	RT @Fla_Pol: Don't call it a 'spaceship': New ...	1	0	Twitter Web Client	2019-03-13 16:48:56	272067125	JanelleIrwinFL	Janelle Irwi

Sentiment Function

- Automatically parses tweet text
- Returns two numbers
    - Sentiment Polarity
    - Sentiment Subjectivity

```
def get_sentiment(text):
    tb = TextBlob(text)
    return(tb.sentiment)
```

Polarity: 0 being perfectly neutral, positive being good, negative being bad. Maximum range

-3 <= Polarity <= 3

Subjectivity: How subjective a tweet is

0 being 100% factual and 1 being 100% subjective

Creating a training data set

- Manually labeling data
- Sampled 6 different companies
    - Youtube
    - Apple
    - Tesla
    - Florida Polytechnic University
    - Wells Fargo
    - Facebook

user_id	status_id	text	Total	Diego	Jackie	Richard	Luke	Kahlil
33152005	1105291678422`	I just tried posting @zerohedge's story about @facebook banning the well known financially heterodox @zerohedge. I figured that this was insane. Facebook bans Zerohedge? C'mon. Ladies and Gentleman, Zerohedge was correct: the posts are 'spam' according to Facebook. Stunned. https://t.co/HBVRnuLTAd	#N/A	Positive	Positive	Negative	Negative	Negative
832978972602`	110547790817`	Hey @facebook why is smearing period blood on your face ok, but an ANTI-ABORTION video is considered "graphic" content?! @UnplannedMovie #babylivesmatter #prolife https://t.co/avV0IWtPI1				Negative		Negative
15976705	110526379895`	Hold on, that not fair or good business @facebook. Seeking to silence criticism of your business practices is something that happens in an authoritarian state,				Negative		Negative
181096072	110582632797`	@Twitter is too dope even @facebook @instagram @Snapchat @Tinder etc... have twitter accounts 🤖				Neutral		Neutral
356758491	110582661980`	@facebook I am totally shocked that Facebook had access to mypaypal. You have charged me for something I didn't want or agree to pay for. Faceless organ				Negative		Neutral

Future Work

- Complete labeling training set
- Feature engineering on training set
    - Add sentiment polarity and subjectivity
    - N-gram analysis
    - Tf, tf-idf, idf
- Train a classification model on the training set
- Allow for live tweets to be automatically classified as one of the labels
    - (Positive, Negative, Neutral, Issue)
- Possibly add a GUI

Challenges

- Tweets need to be cleaned and sorted appropriately
    - Spam
    - Media
    - Retweets
    - Non-english
- Labeling the training set is cumbersome
- Rate limit on API
- Data is constantly changing

Questions?