Topic - Language analysis of tweets related to mental health on Twitter
Data Source - Twitter API, collected by using keywords and hashtags
Analysis Types - Clustering, Natural Language Processing, Regex, Naive Bayes, Word Vectors
Disciplines - Anthropology, Psychology, Linguistics and Data Science
Visualization - definitely
Author - Mo Johnson

Background/Justification for Topic

Twitter and Mental Health Data
This study was the initial study that I read that inspired me to attempt a project that uses twitter data. The authors look
http://www.cs.jhu.edu/~mdredze/publications/2014_icwsm_ptsd.pdf

The following study uses NLP and Language Models to look at Bipolar, PTSD, SAD. They also examine tweet rates, number of mentions and other metrics that are available from twitter data. The authors also examine sentiment, anxiety and anger.
http://www.cs.jhu.edu/~mdredze/publications/2014_acl_mental_health.pdf

This study looks at how depressed and non depressed individuals use language differently on twitter. They discuss LDA models.
http://www.umiacs.umd.edu/~daithang/clpsych2.pdf

Twitter and Health
A University of Pennsylvania study looked at positive versus negative tweets and incidence of mortality due to heart disease: http://www.care2.com/greenliving/what-does-your-twitter-account-say-about-your-health.html

This is one of many projects that look at correlations between word sentiment and meaning and health.

Words and Mental Health
Pennebaker is a psychology professor and research at University of Texas at Austin who has studied the use of words (style, content) and pronouns to analyze a diverse set of topics including depression, suicide prone-ness and social bonding after trauma. He has also studied the use of words (writing narratives) to heal from trauma. I admire his work and would like to use some of his concepts and research to inform some of my inquiry. Almost all the mental health studies have referenced his LWIC analysis, and show how some language models are superior to the LWIC.
His website: http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/home2000/jwphome.htm

Using data science to understand mental health is timely
The director of the National Institute of Mental Health just accepted a job at Google: http://fusion.net/story/221123/thomas-insel-google-alphabet-life-sciences-mental-health/

Data Collection using Twitter

I will use the following twitter fields to analyze the data collected from twitter.

Users

from: https://dev.twitter.com/overview/api/users

name	description	why?
description	String: The user-defined UTF-8 string describing their account.	information about the user
entities	variety: Entities which have been parsed out of the url or description fields defined by the user, includes hashtags.	information about the user
favourites_count	int: The number of tweets this user has favorited in the account's lifetime. British spelling used in the field name for historical reasons.	user activity
followers_count	Int: The number of followers this account currently has.	user reach: who sees what that person posts?
friends_count	Int: The number of users this account is following (AKA their "followings").	user activity
geo_enabled	Boolean: When true, indicates that the user has enabled the possibility of geotagging their Tweets. This field must be true for the current user to attach geographic data	to help with location. Not all users have this enabled, but many users have a self-described location.
id_str	string: The string representation of the unique identifier for this User.	better than id (too large), and identies the user
lang	string: The BCP 47 code for the user's self-declared user interface language. May or may not have anything to do with the content of their Tweets.	should all be in english, based on tweepy, but just in case
listed_count	int: The number of public lists that this user is a member of.	user reach
location	string: The user-defined location for this account's profile.	geo-info on user

name	description	why?
name	string: The name of the user, as they've defined it. Not necessarily a person's name.	user info, considering not using it because of mental health nature of project
screen_name	string: The screen name, handle, or alias that this user identifies themselves with. screen_names are unique but subject to change	handle, but will rely on id_str for unique user id, because these are subject to change
statuses_count	int: The number of tweets (including retweets) issued by the user.	measures user activity - how active they are in the given time period
time_zone		time/geo info
utc_offset		
verified	Boolean: When true, indicates that the user has a verified account.	user info

Tweets
from: https://dev.twitter.com/overview/api/tweets

name	description	why?
coordinates	collection of float: longitude and latitude of where tweet was composed if geotagging is enabled, Represents the geographic location of this Tweet as reported by the user or client application. type: point	sort by location. eg. look at ptsd tweets close to a military base
created_at	string: UTC time when this Tweet was created.	when are tweets about X made? 2 am? 10 am? weekends?
entitites	variety of things including hashtags and url, Entities which have been parsed out of the text of the Tweet.	Which hashtags are associated with other hashtags or other key words collected. Useful links? (not sure how to analyze links)
favorite_count	integrer	how many users agreed with or resonated with the tweet
id_str	string: The string representation of the unique identifier for this Tweet.	to identify unique tweets, especially if the same tweets carry two or more key words
lang	string	double check to make sure its english languagef
places	variety of things from Places: When present, indicates that the tweet is associated (but not necessarily originating from) a	
retweet_count	Int: Number of times this Tweet has been retweeted.	to see the reach of a specific tweet
text	String: The actual UTF-8 text of the status update. See twitter-text for details on what is currently considered valid characters.	the text of the tweet— what I will analyze using the NLP
user	This information is outlined in the chart above.	

Hashtags/ Key Words

The Berkeley Media school has studied the use of hashtags for trauma and childhood trauma and has listed a set of recognized hashtags here: http://www.bmsg.org/resources/publications/talking-about-childhood-trauma-adversity-twitter-overview-hashtags

Mind Magazine has a list of hashtags used to people who are creating a community around a specific mental health experience and stigma felt by community members: http://

www.mind.org.uk/information-support/your-stories/mental-health-hashtags-on-twitter/#.Vi7RsSBViko

Code

for collecting data from twitter and exporting it to a cvs file

```
import tweepy
import csv

NUM_TWEETS = 200


api_key = 'WrDp5GtkmcPN2AgicZ6qpExxC'
api_secret = 'TJZvpBrhZN2uWIGXrLAEUB8vlkUjUY5KGqmnlewBEnTQYC1hm1'
access_token = '55374931-Avn0K2PzESwbaoZ5kkihxZtNwdSxmoAQ1vZuAiBTI'
access_secret = 'KSqUWQSy2MHfX35GqIYTTQlfFVedei5G8Wdu3eAVZeYdq'


auth = tweepy.OAuthHandler(api_key, api_secret)
auth.set_access_token(access_token, access_secret)
api = tweepy.API(auth)

csvFile = open('ptsd_tweets.csv', 'a')
csvWriter = csv.writer(csvFile)

tweets = tweepy.Cursor(api.search,
                       q="ptsd",
                       result_type="recent",
                       lang="en").items(NUM_TWEETS)


for tweet in tweets:
          csvWriter.writerow([tweet.user.id_str, tweet.id_str, tweet.text.encode('utf-8'),
tweet.retweet_count])
          #print tweet.user.id_str, tweet.id_str, tweet.text, tweet.retweet_count


csvFile.close()
```

The preliminary ptsd_tweets.csv is in my github as well. To not overrun API, I do 200 tweets at a time.

Note1 : For the q="ptsd", I am going to run several different queries based on the the Berkeley article above, and I will run tweep twice a week to collect data. A friend who is a psychoanalyst and linguist is helping me chose the keywords based on his experiences.
Note2: To have a "control" file, I will chose some words not related to mental health.

Future Analysis

I plan to use the techniques learned in the NLP lectures to build code and programs to analyze the data. I am working on cleaning the data.