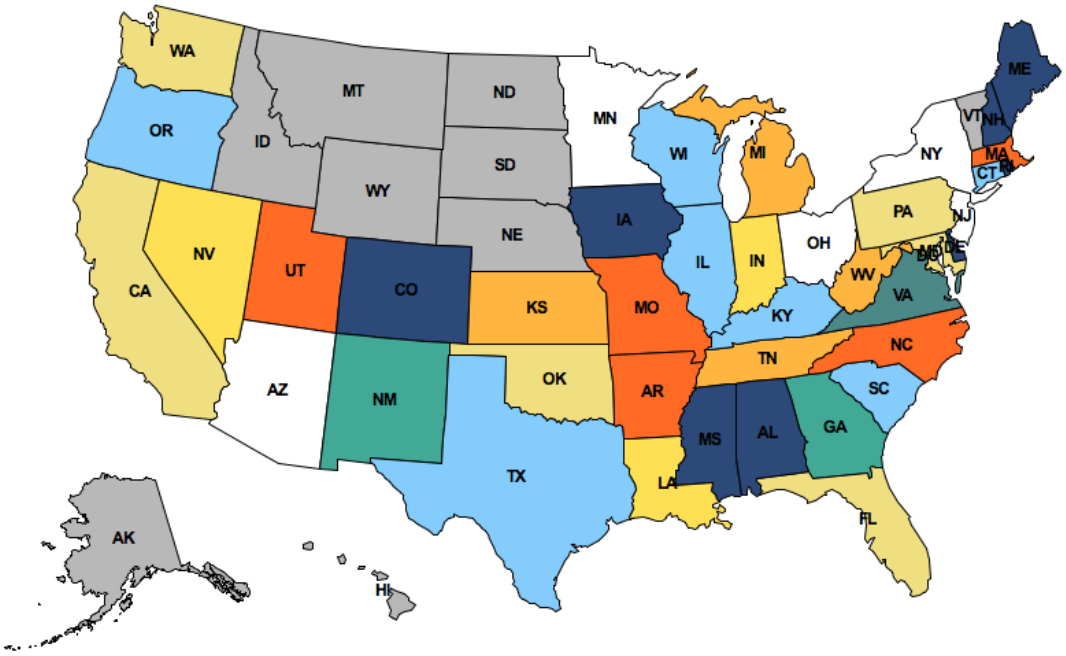
Assignment 2. Twitter Trends

Project objectives

In this project, you will develop a visualization of twitter data across the USA. The map displayed below depicts how the people in different states feel about California.



This image is generated by:

1. Collecting public Twitter posts (Tweets) that have been tagged with geographic locations and filtering for those that contain the «cali» query term.
2. Assigning a sentiment (positive or negative) to each Tweet, based on all of the words it contains.
3. Aggregating Tweets by the state with the closest geographic center, and finally
4. Coloring each state according to the aggregate sentiment of its Tweets. Yellow means positive sentiment; blue means negative.

The details of how to conduct each of these steps is contained within the project description. By the end of this project, you will be able to map the sentiment of any word or phrase.

Input data

The Data directory contains all the data files needed for the project, and it's necessary to run the project. Each file in this folder (cali_tweets2014.txt, family_tweets2014.txt, football_tweets2014.txt, movie_tweets2014.txt, ...) contains collection of tweets on a related topic: California, family, football, movies, ...

The Images directory contains the correct maps that your program should produce by the end of the project for the given terms.

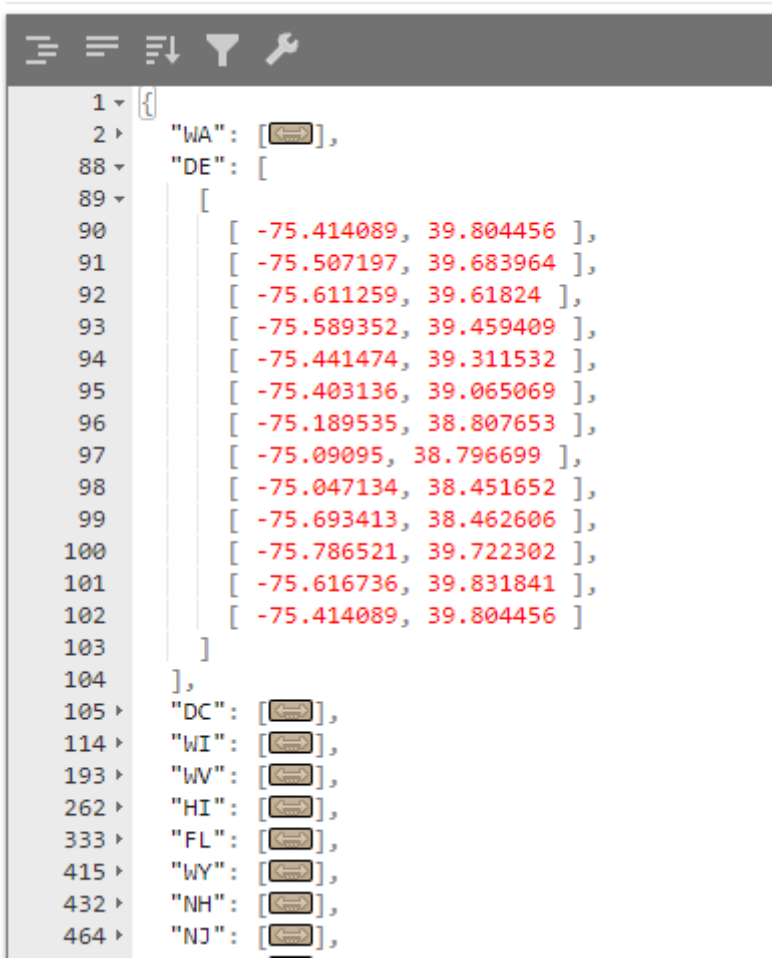
Every tweet is represented as following: the latitude of the tweet's location, the longitude of the tweet's location, date and time when the tweet was posted, the text of the tweet.

```
[31.2784485, -90.46151438]  _     2014-02-16 03:14:35    watching a
football life
[41.68770968, -71.16827271] _     2014-02-16 03:15:18    Is it bad
that I already miss football season?
[37.96682961, -84.45854165] _     2014-02-16 03:15:57    but a few
years ago FL had a good football team and now they have a good
basketball team that's not fair
[40.092657, -88.23893792]   _     2014-02-16 03:16:41    It's
really sad when the ILL vs OSU football game has more points than
the basketball game. #pathetic
[26.70290653, -80.22700978] _     2014-02-16 03:16:47
      @KiCKN_KLOUDS: "@CodyCo3HUNNA: All I Need Is My 2 F's=Family
&amp; Football I Swear" &amp; That Rider during The season lol
[32.52288415, -84.87604275] _     2014-02-16 03:18:32    Man I miss
football already ....Basketball just don't excite me if nobody
getting dunked on or crossed up
[34.05222357, -83.39203041] _     2014-02-16 03:18:40    @An_dre25
no doubt lol. And also an hour of football b4 the workout!
```

Some words of the tweet are associated with positive or negative sentiment, but most are not. The sentiment of some individual words can be found in the Data/sentiments.csv text file.

```
a lot,0.25
a priori,0.25
a-one,0.625
abandon,-0.375
abandoned person,-0.375
abandoned,-0.375
abase,-0.625
abash,-0.25
abashed,-0.375
abashment,-0.375
abasia trepidans,-0.375
abasia,-0.5
abasic,-0.25
abatable,0.625
abatic,-0.25
abaxially,0.25
abbot,0.5
abdias,-0.25
abdicable,0.375
abdominal breathing,-0.25
aberrant,0.375
aberrate,-0.25
abetalipoproteinemia,-0.75
abhor,-0.25
```

Json file Data/states.json contains data on geographical location of each U.S. state. Each state is keyed by its two-letter postal code (WA, FL, HI) and the shape. The shape of a state is represented as a list of polygons. Some states (e.g. Hawaii) consist of multiple polygons, but most states (e.g. Colorado) consist of only one polygon (represented as a length-one list of polygons).



Phase 1: The Feelings in Tweets

In this phase, you have to split the text of a tweet into words, and calculate the amount of positive or negative feeling in a tweet.

Problem 1. Before we can analyze the feelings in tweets, we need to access its words. You have to extract words from tweet. Assume that a word is any consecutive substring of text that consists only of letters.

Problem 2. Analyze tweet sentiment: take a tweet and return a single number averaging the weights of sentiment-carrying words in the tweet, or None if none of the words in the tweet carry a sentiment weight.

Phase 2: The Mood of the Nation

In this phase, you will determine the state that a tweet is coming from, group tweets by state, and calculate the average positive or negative feeling in all the tweets associated with a state.

Problem 3. Implement function which returns the two-letter postal code of the state that is closest to the location of a tweet.

Problem 4. Group tweets by state. Take a list of all tweets and return a dictionary. The keys of the returned dictionary are state postal codes, and the values are lists of tweets that appear closer to that state's center than any other.

Problem 5. Implement calculate_average_sentiments. This function takes the dictionary returned by group_tweets_by_state and also returns a dictionary. The keys of the returned dictionary are the state names (two-letter postal codes), and the values are average sentiment values for all the tweets in that state.

If a state has no tweets with sentiment values, leave it out of the dictionary entirely. Do not include states with no tweets, or with tweets that have no sentiment, with a zero sentiment value. Zero represents neutral sentiment, not unknown sentiment. States with unknown sentiment will appear gray, while states with neutral sentiment will appear white.

Problem 6. You should now be able to draw maps that are colored by sentiment corresponding to tweets that contain a given term. The correct map for «cali» appears at the top of this document.

Original version of the project description (from University of California, Berkeley) you can find here https://inst.eecs.berkeley.edu/~cs61a/fa14/proj/trends/