Sentiment in San Francisco

Jared Wilber SID #24881068

December 13, 2016

Abstract

Measuring public sentiment is important to policymakers, private-industry, and researchers alike. The recent exponential growth in publically available social media data allows for more geographic specific sentiment analysis than ever before. In this paper geo-tagged data from Twitter will be used for several purposes. It will be classified on a scale of -1 to 1 with regards to sentiment (1: positive, -1: negative). It will be used to generate a sentiment map of San Francisco. Finally, it will be used to determine that public sentiment changes with proximity to public parks. For reproducibility, everything is available on github.com/jwilber.

Introduction

In 2006, Twitter was created as a microblogging site. Today it is used by over 500 million people . As a dataset, Twitter has proved invaluable to researchers and has been utilized for a number of tasks, such as predicting financial markets, political affiliation, and analyzing the after-effects of natural disasters. It is also used for sentiment analysis, with results yielding similar resuls to traditional metrics, such as polling. In what follows, we'll combine standard sentiment analysis techniques with geography in an analysis to determine that public sentiment is different near public parks.

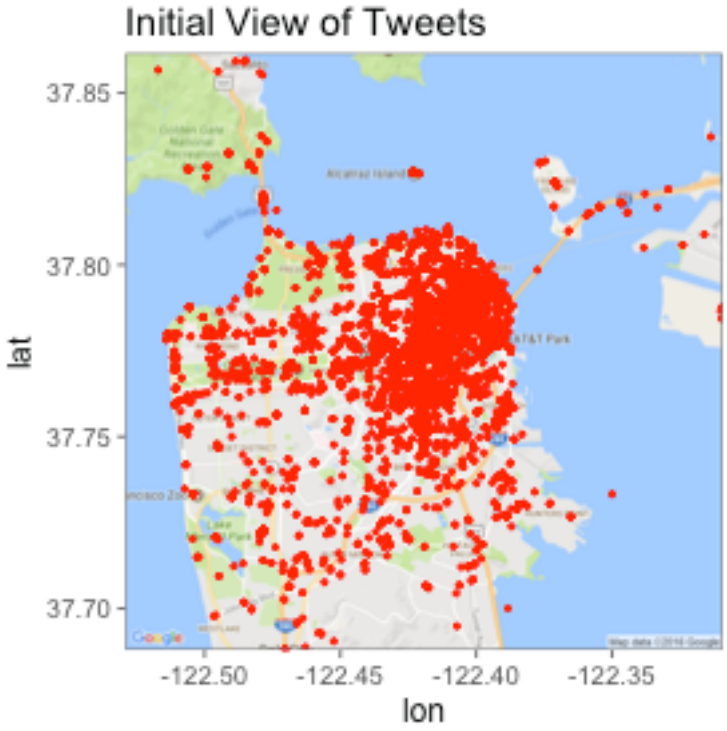
Data

As stated, the data for the analysis comes from Twitter. It comes in two forms. First, a dataset was obtained [http://www.followthehashtag.com/datasets/free-twitter-dataset-usa-200000-free-usa-tweets/] with 200,000 geo-tagged tweets from the years 2014 and 2015 in the United States. This dataset provides utility in that it prevents our data from being too confounded with recent events.

Second, approximately 1 million tweets were scraped via the Twitter API through R into a mongodb database. These scraped tweets were limited to the greater San Francisco area. The two datasets were then merged. The final dataset was restricted to only those tweets that were geotagged and free of "noise" hashtags (such as #werehiring or #weatherupdate), as these just distort the average sentiment in the data.

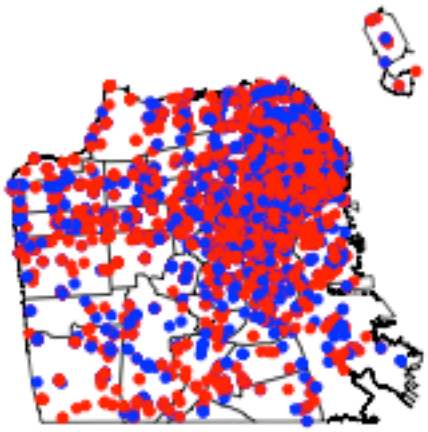
The final dataset consists of roughly 13,000 tweets.

Below is an initial view of our data.



[h]

Following this, the sentiment of the tweets was classified. This sentiment was classified using standard NLP techniques; namely, a continuous bag of words model run through a gradient-boosted model. From this classification, two sentiment features were created, one showing the continuous sentiment score in the range -1,1, and another showing the binary output (positive/negative). The binary output was created from the continuous output, with scores greater than 0 assigned as positive, lesser as negative.



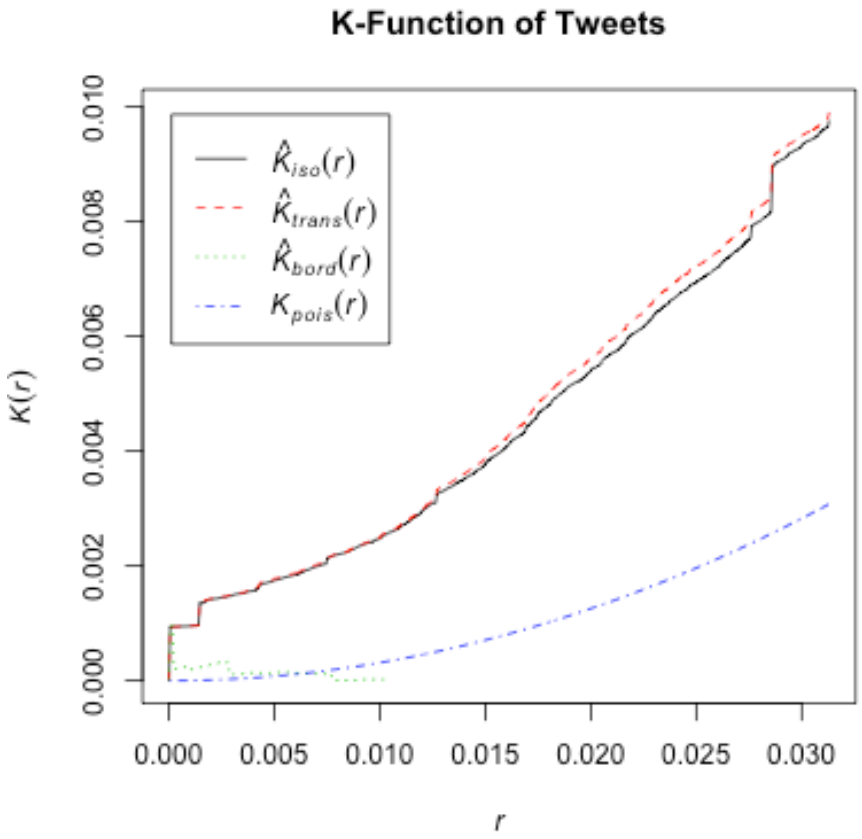
[h]

At first glance, the spatial distribution of sentiment appears unclustered, with positive and negative values appearing in a seemingly haphazard manner.

Analysis

The scope of our analysis is limited to San Francisco. Thus, all tweets not in San Francisco were eliminated. This was achieved via map algebra: a mask was created over our shapefile of San Francisco, with all tweets outside that mask eliminated. As a first step in analysis, we analyze the data using a Poisson-point-process technique known as the K function. The K function evaluates the spatial distribution of our points in relation to complete spatial randomness (CSR).

We view our K function plot.

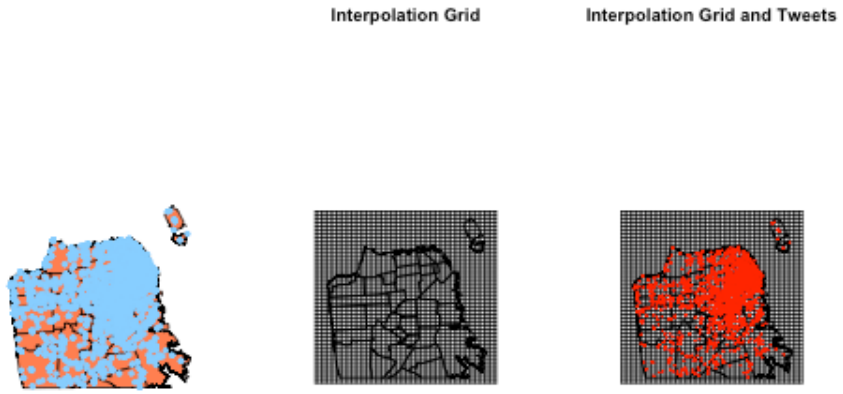


[h]

In our case, our data is above the we can see that the K function yields results suggesting strong clustering in our data. People that use Twitter probably share similar characteristics, and thus are likely clustered spatially. Furthermore, certain locations receive a much larger influx of individuals than do others, so our K function results are expected.

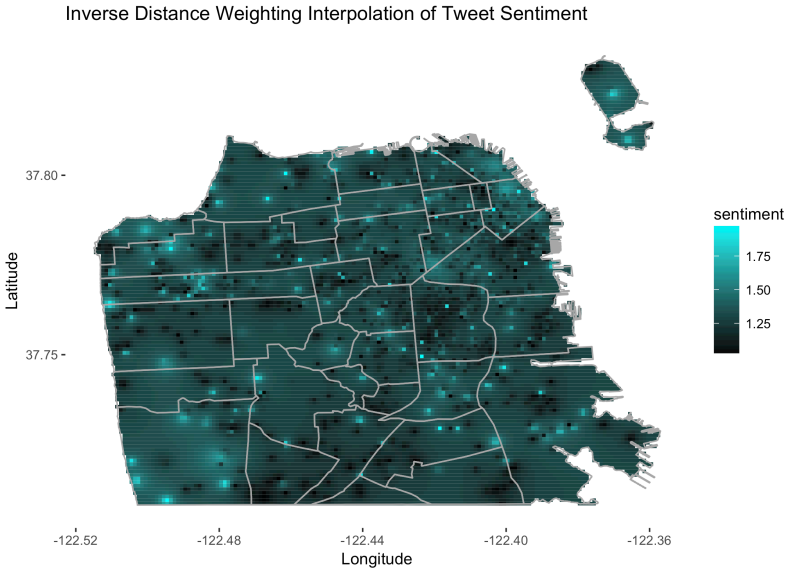
Our tweets are clustered as expected. How is sentiment distributed spatially? Recall our data isn't exhaustive of the total space in San Francisco. To completely cover it in tweets will require a lot more tweets, some in areas where people seldom go. This could require months of scraping, and so is not practical. Thus, to determine spatial distribution of sentiment, we'll employ interpolation techniques.

First, the data is split up into very small grids of size .001 latitude and longitude. Then, we use inverse-distance weighting to estimate the sentiment of each grid. The outline for this process is shown below:



[h]

Inverse distance weighting is a deterministive method of interpolating a set of scattered points. Essentially, the value of each grid will be calculated as a weighted average of all other points, with more weight being assigned to those points that are closer. In this manner we create an interpolated plot of sentiments for all of San Francisco.



[h]

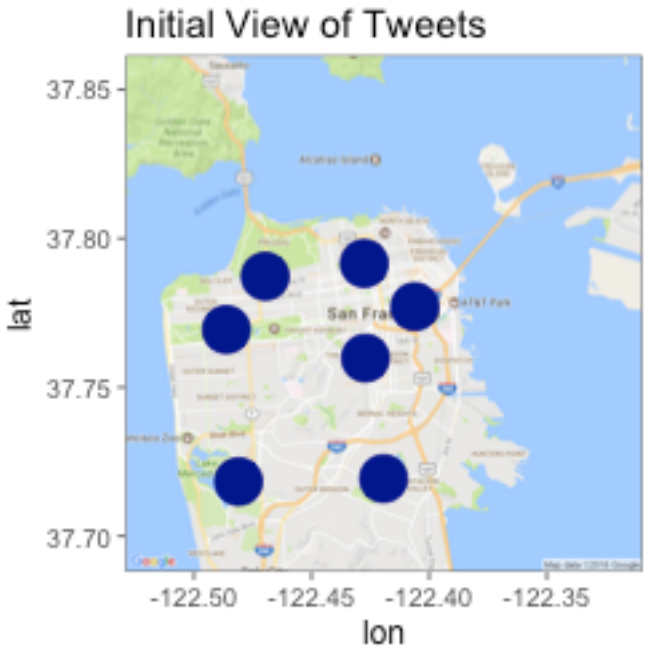
The plot reveals interesting patterns. We can see that the South-West portion of San Francisco appears much happier. This is in accordance with our hypothesis that people are happier near public parks, as that particular region offers Lake Merced Park, Park Merced, Pine Lake Park, and Sigmund Stern Recreation Grove. We can also see that the North-East region of San Francisco appears to be less happy, particularly in the Financial District. It's worth noting that this contrast may be a function of the Financial District having more tweets, and therefore a higher chance of negative tweets. We'll investigate this analytically soon. Note that a significant chunk of the map is colored a shade between black and cyan; this reflects a shortage of data. Were we to scrape data for a couple months or a year, we'd expect to see more fine-grained patterns in the sentiment. Still, our map does quite a good job of mapping sentiment with less than one month's worth of scraped data.

Results

The sentiment reveals spatial trends in public sentiment. Moreover, it reveals trends that we anticipated. But are these trends statistically significantly? To investigate, we'll perform a statistical test. We'll treat each tweet as a unit belonging to an observational study where-in which the treatment is regarded as whether or not a tweet among the closest ten to a public park. We use the following public parks.

Park	Longitude	Latitude
Dolores Park	-122.4271	37.7598
Golden Gate Park	-122.4862	37.7694
Lafayette Park	-122.4276	37.7916
Park Merced	-122.4810	37.7183
John Mclaren Park	-122.4194	37.7193
Victoria Manalo Draves Park	-122.4061	37.7771
Mountain Lake Park	-122.4697	37.7873

The parks were chosen so as not to be clustered near eachother.



[h]

As the plot shows, they encompass most regions of San Francisco in a concentric manner, which should yield a more representative outcome.

In this way, our data is divided up into two groups, one for each treatment level (among top ten closest tweets to public park or not).

Let β represent the difference in mean sentiment between the two treatment groups. Then we calibrate our test in the following manner:

$$H_0 : \beta_{=}0$$

$$H_1 : \beta \neq 0$$

In other words, our null hypothesis dictates that individuals near public parks have the same sentiment as those far away, implying public parks don't make people happier or angrier.

We'll carry this out via a permutation test. A permutation test is a nonparametric method of statistical inference that tests a specific null hypothesis that the treatment levels we are comparing are completely equivalent and serve only as labels; i.e. that the responses we observed for our units would be the same no matter which treatments had been applied. It belongs in the family of resampling methods, much like monte carlo simulation.

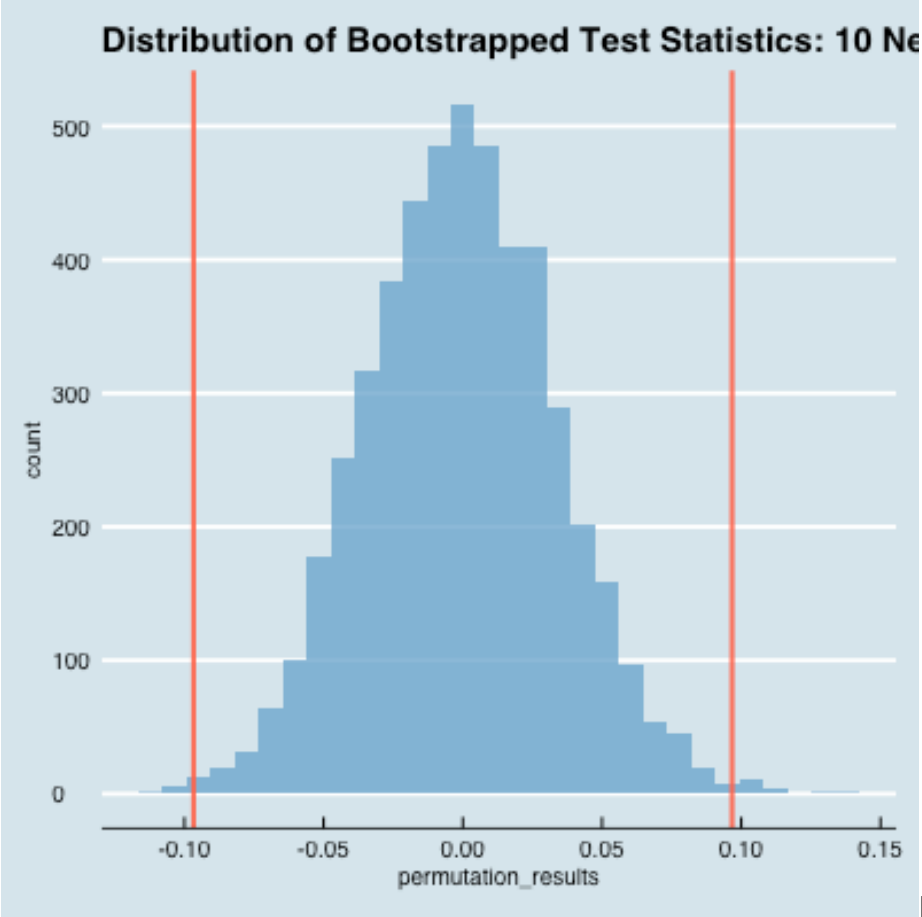
We use this as our particular choice of test for a number of reasons. First off, our observational data fails a number of statistical assumptions underlying traditional tests, such as t-tests or ANOVA. Our data is extremely unbalanced: the ten closest tweets consists of 70 tweets total, while the other level holds over 10,000 tweets. Furthermore, the residuals of our data don't show heteroskedasticity. A permutation test avoids these complications.

We proceed as follows. First, we get our initial test statistic: the difference in mean sentiment between our two different treatment groups.

Next, we permute the treatments among our tweets, then calculate the same test statistic on the newly shuffled data.

We continue in this manner, permuting the treatment labels and recalculating test statistics 10,000 times, building a (Gaussian) distribution of test statistics.

Finally, we view where our original test stastic exists in this distribution.



[h]

The above image details our test statistic distribution. The red vertical lines

display where our original test statistic belongs in the distribution. This corresponds to a p-value of **0.0048**.

P-Value	0.0048

In other words, observing the spatial distribution of sentiment we did, assuming sentiment was not different with regard to proximity to public parks (our null hypothesis), has a probability of less than 1 percent.

This is clearly a significant result, and we can conclude that people have different sentiment near public parks. As for the direction of that sentiment, we can infer from our previous interpolation map that people are happier. On the one hand, obtaining a significant result is very exciting. However, this result is hardly unexpected; people don't just end up in public parks, they go their by choice. Moreover, if they're willing to drag themselves to a public park, they're probably doing so because it makes them happy. Numerous studies have also concluded that public parks result in happier citizens.

Conclusions

In this paper, we scraped tweets, classified their sentiment, and analysed them spatially. A sentiment map of San Francisco was constructed and it was determined that people near public parks were happier. This research supports demand for governments to construct and maintain green areas in their cities. It also supports the broader notion that people may benefit from spending more time outdoors in public parks. While it's not clear what the far-reaching benefits of these suggestions are, it's clear that people will be happier. That said, it's possible that confounding is occurring. For example, is it the case that people are happier near public parks, or that people who use twitter in public parks are happier? Regardless, the results are interesting and line up with general expectation. Further analysis will require more time scraping data.