Spatio Sentiment Analysis of NBA Playoffs 2018 Based on Twitter

By Xuefei Liu, Chichiger Shyy, Yuxiao Wang

Introduction

This year, the NBA Playoffs have involved over millions of fans using Twitter to talk about the games and teams they are rooting for. Such a popular and well promoted event such as the annual NBA All-Star Game is worth analyzing about. With the playoffs lasting from the end of April until the first week of June, there will be a large amount of raw data from Twitter to use.

The aim of this study is to investigate the association between different states and different sports team affections. We plan to create state level indicator of team affection. For simplicity, only two choices are considered: East and West. Based on the team that they are talking about and which division the team is part of, we will assign each user to East or West. For example, the Celtics, Cavs, and Raptors will be part of the East while the Warriors, Rockets, and Spurs are part of the West. We would like to find the relation or factors in whether local citizens of a state favor their home team or not. Data from social media will help to uncover the team/player preference in certain area, especially Twitter. We would use Twitter as our primary source and restrict to geographic tweets. All the tweets about certain players and teams would be counted toward their divisions. We would use several existing spatio-temporal methods, such as compass, to evaluate the support in different states.

Collecting Data

Twitter API RESTful provided us with each tweet in JSON format, and the data includes information about the attributes. We wanted to restrict ourselves to geo-tagged tweets from the USA, which means that we needed to filter and get all the tweets that have the Place ID that is included inside the United States. We fetched the data with a query that had a specific USA ID (provided by Twitter).

After we created the app on Twitter Developer System and have the keys for the Twitter API, we used Tweepy, an open source library that gave us access to the Twitter API. we got each day's tweets using the function 'findTweetsMoreTags', which executed multiple queries with different keywords/hashtags and place id of USA. Then we stored tweets for each day in a MongoDB database. After we have each day's tweets data in different database, we will need to combine these database into one so that we can do split function with the State easily. To do this, inside the 'dataAnalyze.py', we use function 'dataCombine' to combine all tweets from different days into one database, along with filtering the Tweet unique ID to make sure there is no Tweet added multiple times.

Organizing Data

There are many limitations and things we can improve in this project. Looking for a useful dataset is the biggest limitation we had. We take 5 factors in our consideration, because these datasets are accessible. Even though we find a dataset, some of data are not in a good quality. For example, in the average rent price dataset from Zillow, there are many blanks in certain zip code. To fill up the blank, our team calculate the average rent price in Boston and use the rent to run the tests. In the future, we should talk more factors in our calculation.

LSI and SVD

We used constraint satisfaction to get the best zip code in the Boston area. For each factor we considered, we calculated the mean μ and standard deviation σ . We only took valid data form [μ -3σ, μ + 3σ]. We calculated the correlation between factors and then we used correlations to calculate weight for each factors. For each zip code, we multiplied weights to each factors and got the final evaluation score for each area. We then used greedy algorithm to get the most valuable areas to invest.

Conclusion

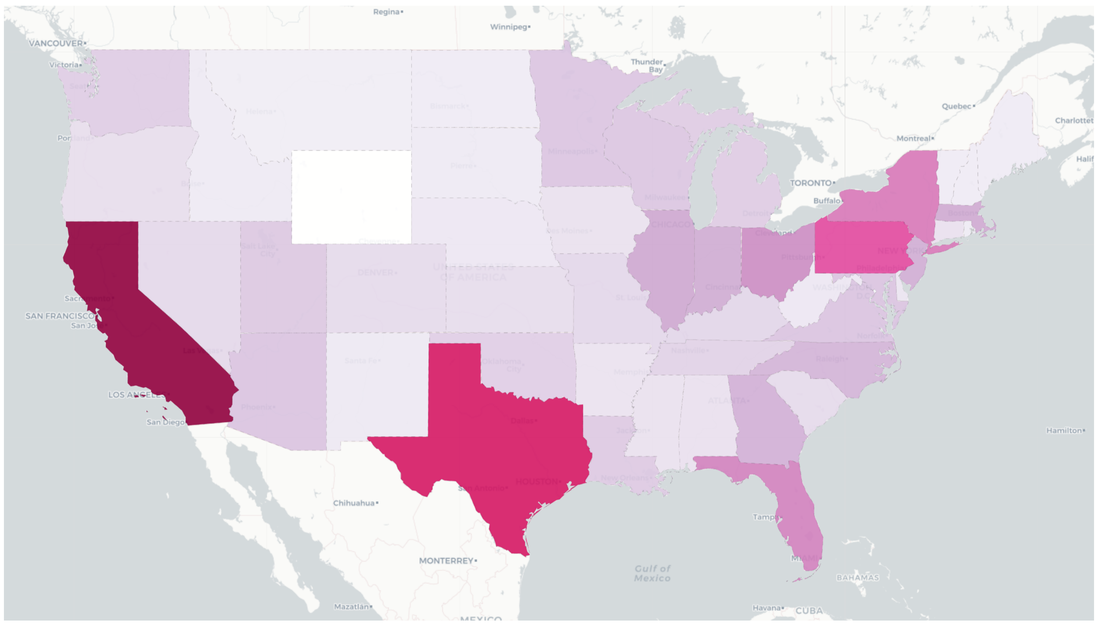

Figure 1: Map of Boston showing which zip code areas are most valuable to invest with respect to our calculating scores by facility.

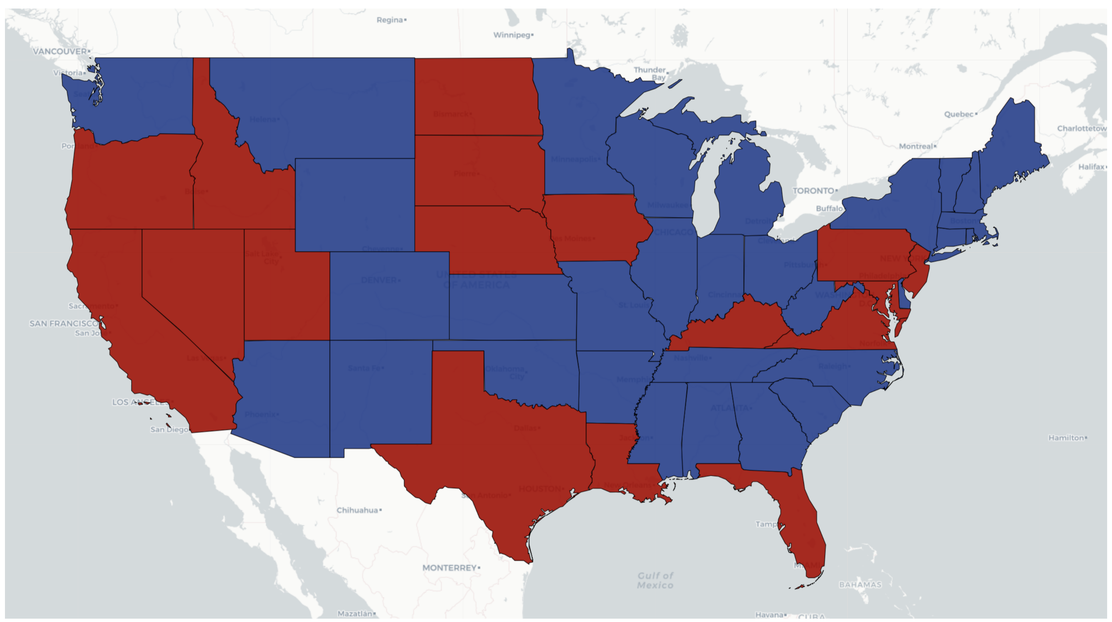

Figure 2: Table showing correlation coefficients and weights