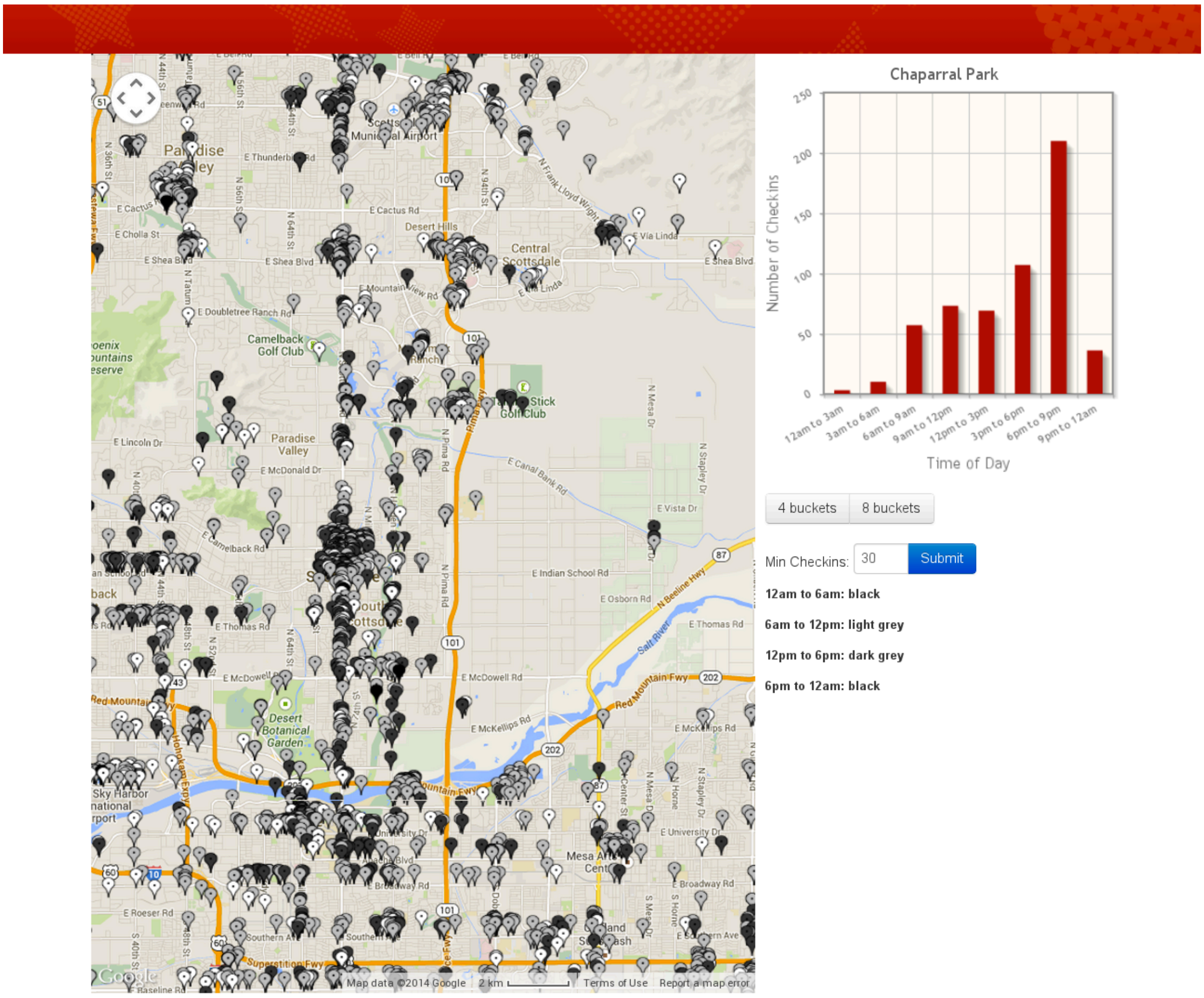
Ismail Khan
CSCI 1951A

Predicting and Visualizing Check-In Distributions for Yelp Businesses



12am to 6am: black

6am to 12pm: light grey

12pm to 6pm: dark grey

6pm to 12am: black

Yelp has generously released information regarding 15,000+ businesses in the greater Phoenix, Arizona metropolitan area as part of the Yelp Dataset Challenge. The goals of this project were twofold:

The first goal was to use the text of user-submitted reviews for each business to build a classifier that predicts the time of day that a business is busiest – that is, the time of day that a business has the greatest number of Yelp check-ins. We used the frequency of each word in the reviews of each business as "features", and split one day into four six-hour buckets (to simplify the classification process). We limited the number of features to 1000 features using three different feature selection methods combined with two different classification methods. The best overall classifier turned out to be a one-versus-rest logistic regression classifier coupled with chi-squared feature selection. On our test dataset, this classifier gave us an accuracy of 68.74%.

The second goal was to build a web-based visualization tool to easily view the check-in distribution of these businesses. The tool is built using a Node.js server, SQLite database, the Google Maps API for Javascript, and the JQPlot plotting library. Each marker represents one business and is color-coded by the time of day that is has the greatest number of check-ins. Upon clicking on a marker, a plot of the check-in distribution for a single business is displayed. Users can change the color gradient from 4 to 8 different colors, and can modify the minimum number of check-ins needed for a business to be displayed on the map. The visualization tool is displayed on first page of this abstract.