TP-D1: Team 7

Title: Optimizing Businesses for Localized Climates

Owen Davidson
Dylan Trinkner
Jake Marrapode

• Datasets

https://www.yelp.com/dataset/documentation/main

This is a comprehensive set of Yelp business reviews submitted by users on Yelp.com. The set contains 6.6 million reviews for 192,609 businesses. Records include unique business identifiers, location data, total reviews, individual review text, business hours, and Yelp-defined business categories for classification.

https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/NORMAL_DLY_documentation.pdf

This is a huge set of normalized federal weather data organized by hour, day, month, season, and year from 1981-2010. Records include normalized average temperature in several different bases, location data such at latitude and longitude, recorded precipitation, predicted precipitation, wind, and dew point, among other extraneous parameters.

• Problem Characterization

For our project, we want to analyze the relationship between various business types and climate. Specifically, our goal is to identify whether or not a theoretical optimized business can be created to maximize customer satisfaction within a specific climate, and additionally identify if an optimized business that ensures customer satisfaction regardless of climate exists. To accomplish this, we will need to determine if a business's local climate affect perceived customer satisfaction by analyzing reviews of business' in similar categories across multiple climates.

• Currently Published Related Work

In their paper titled *Urban Mobility Sensing Analysis through a Layered Sensing Approach ,* K. Machado and T. H. Silva use the Yelp dataset as well as weather data on humidity and temperature among others. They used yelp "check-ins" with weather data and to find the relationship between weather and consumers going to restaurants.

In their paper titled *Improving Restaurants by Extracting Subtopics from Yelp Reviews,* James Huang, Stephanie Rogers, and Eunkwang Joo use the Yelp dataset to describe latent subtopics in restaurant reviews. This is accomplished using an online Latent Dirichlet Allocation (LDA) algorithm, which is a generative probabilistic model. The end result of their work provided a method to evaluate restaurants based on a topic that contributes to the overall rating but does not have an individual rating, and the group was also able to leverage this to perform a temporal analysis of restaurants to see which restaurants were rated better at specific times.

• Analytic Tasks

We will use the NOAA weather set to create a subset of biomes with similar weather properties. Using this subset, we can categorize every business tracked by Yelp into one of these biomes, creating a sampling set. We can then create an ‘average’ business for each biome to be used as a sample mean. Then, using ANOVA (Analysis of Variance) testing and Bonferroni corrections (or similar), we can determine which of these sample means is significantly different from any others. Samples that are not significantly different from each other can later be combined to create a theoretically more accurate ‘optimal’ business for biomes in which climate does not significantly affect business success.

• Evaluating Effectiveness

ANOVA F-testing is generally sound, but by default it increases error rates in determining whether or not sample means are indeed different. To this end, we will start by using Bonferroni corrections to reduce this error rate. If this fails, i.e. the probability that sample means differ or remain the same is repeatedly erroneous or implausible, we will attempt to use different correction mechanisms to reduce this error rate. In short, we will be closely monitoring Group and Error squared means to detect and avoid miscalculation and inflated error rates.

In order to compare business metadata, we will have to implement our own way of reliably comparing and averaging the values we receive. Because of this, it may be necessary to create our own correction mechanism(s) that normalize the data accordingly. The form this may take, however, remains to be seen.

Runtime and throughput will also be a concern when implementing the source code for this project. Keeping runs within a reasonable timeframe is crucial. We will create a baseline metric for simply reading and writing our desired data, and work to ensure our final product does not unreasonably exceed this.

Similar Study References

J. Huang, S. Rogers, and E. Joo, “Improving Restaurants by Extracting Subtopics from Yelp Reviews,” iConference 2014 (Social Media Expo).

K. Machado, T. H. Silva, P. O. V. D. Melo, E. Cerqueira, and A. A. Loureiro, “Urban Mobility Sensing Analysis through a Layered Sensing Approach,” *2015 IEEE International Conference on Mobile Services*, 2015.