

Final Project Report

I. Motivation

After several years of successful operation, Yelp Inc. has amassed an enormous amount of raw data on all kinds of business. In this project, we are particularly interested in doing the text mining on customers' reviews in order to infer the future business attention and trend. For the time limit, we focus our study in Arizona State.

II. Data

The Data we used is from Yelp.inc, we used the business table and review table. There were some issues when we loaded the data into oracle especially with review table. Some review text is longer than 4000-string length so we just skipped those very long reviews. Finally we loaded 30K records into oracle. After that, we did many table transformations to create the review table for several states and loaded them into R. We finally generated four review tables for four states. All these data is in the data directory.

We did many table transformation in SQL Developer.

III. Description

Code Description: Our code can be broken into three parts. First part is to create a bubble plot and an interactive plot for each state and its average rating. Second part is to create the word clouds for several states. The last part is to create the dendrogram for two states, Arizona and Nevada.

We created a HTML file with all the code and descriptions for users to learn and understand what we are doing in this project. The HTML file is called project.html in the documentation directory.

III. Discussion and Analysis

We provide the analysis details in project.html and our code directory.

From our Analysis, we infer that:

I. Overall or on AVG, pizza are more popular than chicken in AZ.

II. Overall or on AVG, pizza-related businesses have better customer service than chicken-related business or restaurants

III. There are going to be a big growth in the number of pizza and chicken restaurants in AZ in the future