Big Data

Visualization Practical Work

Interactive visualization of the yelp dataset

Authors:
Fernando Díaz González
Giorgio Ruffa

January 20, 2018

Contents

1 Introduction

Between all the dataset provided we decided to use the yelp visualization challenge dataset, that can be downloaded here https://www.yelp.com/dataset/challenge in the json format. In particular we used the *business.json* and *review.json* files.

2 Technical Notes

Given the size of the dataset we provide only a sample of it. In case a different sample is needed, attention must be payed as the business sampled must be the same in the two files.

Please be sure to have the *igraph* library at its latest stable version (v1.1.2) otherwise the function "strenght" will be missing.

You can access the online version of the app at: https://fediazgon.shinyapps.io/yelp-business-viz/ (sample data).

3 Problem Characterization in the Application Domain

The domain in which we are operating is the one of traditional business-customer interaction. The yelp dataset offers a wide range of information on the topic, like for example the reviews given by customers to the business or when a user did a "check-in". The dataset also contain a very diversified range of categories a business can belong to. Remember that each business can belong to many categories.

3.1 User Questions

Which kind of domain specific questions the user wants to answer?

1. Which is the relationship between the most common business categories?

2. In which time of the day the customer check-in? Does it coincides with the opening hours?

3. Is it true that different categories have different opening hours? Like bars opens until late and restaurants close sooner.

4. How the business are geographically distributed?

5. Is there a relationship between the number of check-in and the number of reviews?

6. What is the distribution of review score on a determined geography? Are there areas that are more "picky"?

3.2 The Dataset

The dataset offers a rich collection of information, the full documentation can be found here `https://www.yelp.com/dataset/documentation/json`.

In our case we think that the promising values that can be used to answer the user questions are the followings.

In the *business.json* file:

- latitude: float latitude
- longitude: float longitude
- starts: float, star rating, rounded to half-stars
- review_count: interger, number of reviews
- categories: an array of strings of business categories
- hours: an object of key day to value hours, hours are using a 24hr clock

In the *checkin.json* file:

- time: nested object of the day of the week with key of the hour (using a 24hr clock) with the count of check-ins for that hour (e.g. 14:00 - 14:59).

4 Data and Tasks Abstraction

4.1 Action Use

The actions the user need to perform are all included in the *consume* category.

More specifically the user wants to *discover* features of the dataset like if there is a relationship between the business categories, to see if there is any pattern in the check-in behaviour and to look for a relationship between check-in and number of reviews.

The user also wants to *verify* certain hypothesis, like if the opening hour is related to the business category.

Finally we also want to *present* the geographical location of the businesses.

4.2 Action Search

In the case of the geographic business distribution the user just wants to *explore* the location of the business and find if there are some areas that are more or less incline to score high a business.

In the case of the relationship between check-in and number of reviews the user knows exactly what he is looking for and where to look for it. So we are talking about a *lookup* search.

In the remaining cases the user is looking for characteristics in different categories, so in this case the search will be to *browse* between the different categories.

4.3 Action Queries

The user will *compare* check-in per hour and business opening hours. The user will also *compare* number of check-in and number of reviews.

The user want also to *identify* the distribution of reviews in an area (which is a single target).

And finally he wants to *summarize* the relationship between the various categories

4.4 Targets

To understand the relationship between the categories our idea is to have a network representation as a target. Another target would be the correlations for number of check-in and number of reviews. Similarities is a target for the relationship between check-in hours and opening hours. Finally the distribution of reviews is a target involving only one attribute.

5 Interaction and Visual Encoding

In the case of the relationship between categories (user question 1) we have already decided in the previous layer to represent them as a network. Our idea is that the network will have as nodes the names of the categories and a link every time two categories appears in the same business. The weight associated to link will increase each time the link appears in the dataset. Being inspired by the "miserables co-occurrence" adiacency matrix (https://bost.ocks.org/mike/miserables/), we decided to use the same idiom. The user will be able to manipulate the view by selecting the sorting criteria of the matrix by name of category, frequency and finally cluster.

For investigating the similarities between check-in hours and opening hours we decided to apply some aggregation building two matrix with day of the week and hour as rows and columns index. We calculated a scalar field where each cell rapresent the number of check-in in that hour/day and the number of business open in that hour/day. The user have the capability to filter by selecting the category to analyze. This idiom is intended to answer user's questions 2 and 3. About the colormap that we decided to use, we where in search of a purely sequential colormap, without a central reference. For this reason we decided to exploit the R package "viridis" and select the magma colormap. The user will also be able to apply smoothing to the heat map to ease the detection of particular patterns.

We decided to answer the last three question using a single view with multiple coordinated idioms (faceting). The question regarding the geographical distribution of the businesses (the number 4) is addressed using choroplet map where the quantities are encoded as two-dimensional bubbles. The user can select if the size and/or the color of the marker encode the average stars (review score) or the number of reviews (in logarithmic scale!). Interaction in this idiom is performed trough bi-dimensional navigation, by means of zooming and panning (no rotation). This interaction affects also the two others coordinated idioms listed below by selecting only the data that appears on the map. The user can also set the range of review score to examine, this is intended to answer user's question 6. Also to help on question 6, we decided to add a coordinated view as an histogram to show the distribution of the review score. Finally, since we have decided that question 5 will have as a target a correlation between two variables, we decided to add a scatter plot between the logarithm of the number of reviews and the number of check-in. The choice of the colormap for the markers was, this time, more complicated as the color can be associated with two different variables with very different ranges. In fact the number of average stars can only take integer or semi-integer values between 1 and 5, while the logarithm of the number of reviews is de-facto a continuos variable. Both the values does not have a central reference, so we decided to use the same colormap for both. This also to avoid rough chromatic changes from the perspective of the user. The final selected colormap is the "viridis" colormap from the "viridis" package as it works decently both with discrete and continuous values.

6 Algorithmic Implementation

The number of different categories is too high to be represented decently on the adjacency matrix, for this reason we used only the 100 most frequent categories. The cluster detection is done using the "igraph" package and the clusters are associated with the "communities" of the graph.

7 Conclusions

We found this dataset to be incredibly interesting and at the same time Shiny has proven to be a very powerful tool in the development of interactive visualizations. We where amazed by how easy it is to develop this kind of applications even with little experience in the field. We hope that Shiny will prove to be an useful tool in our future.

Speaking about the metodology we have to follow, we must say that after the lectures we where afraid on the mistakes we could, and probably did, make. Data visualition seems to be an easy task to do but it appears to be extremely complicated to do in a proper way.