IDENTIFY LOCAL POPULAR ENTERTAINMENT WITH YELP

Zhuoyu Han
han.zhu@husky.neu.edu

Yaofei Wang
wang.yaof@husky.neu.edu

***Abstract*—When people travel to a new city, they might want to know the most popular entertainment in that place. Our project aims to visualize the local popular entertainment on a map with detailed information. We obtain data source from Yelp, which contains information of thousands of hundreds of business all over the world, like geographic information, customer views and so on. After analyzing whole data set, we visualize results within U.S on a website based on HTML. This website provides the most popular entertainment in each state within U.S., as well as the location and customer rating of this entertainment.**

Keywords—Yelp, entertainment, visualization, HTML, map

I. INTRODUCTION

The motivation of our project is to provide information of best local service for those who travels in another state in US. Though we do have some tools right now to help us, for example, google map and yelp, neither of them is perfect in this case. While google map is good at finding detailed information, it can never decide where to have dinner for you. As to yelp data, it provides very professional review of local service, yet no one uses it as a map tool. Our idea is to combine these two and directly display the best store within the state on map.

II. BACKGROUND

A. Real-world applications

When travelling in another state, people would find our webpage helpful and easy to use. Rather than popping up a bunch of local stores at one time, our website would directly provide people with the best store that is definitely worth a visit within the state, which would potentially save people from deciding where to go.

B. Data Source

We use Yelp dataset obtained from Kaggle [1]. This dataset contains Yelp's businesses, customer reviews, user information, the date of check-in and check-out, the information of searching and customer personalized information.. In total, there are:

- 5,200,000 user reviews
- Information on 174,000 businesses
- The data spans 11 metropolitan areas

For convenience of visualization, we choose a subset of original dataset and the selected features we use are showed as below Chart 1.

feature name	description	data type
business_id	ID of business	string
name	name of business	string
address	address of business	string
city	name of city	string
state	abbreviation of state	string
postal_code	postal code	integer
latitude	latitude	float
longitude	longitude	float
stars	customer rating	float
review_count	count of reviews	integer
is_open	open or closed	0/1
categories	category of business	string

Chart 1. Information of subset

III. EXPERIMENTAL SETUP

A. Dataset analysis.

1) Features: we choose 12 features in this dataset, which containing basic and geography information about the business. More important, the customer rating and count of reviews. The geography information is used for mapping, while are implemented on a map. The basic information containing name, city of business is used for identifying the most popular entertainment within each state. And the customer views are used to select the most popular one.

*2) Labels:*The label is categories which is identified by Yelp. Since there are still many similar business types within oone category, we select only one category as label.

B. Prepare data

*1) Data cleaning:*Data cleaning is the first step for data analysis. Since the original dataset is totally raw which could contain missing data, skewed data or none about features. In our project, we apply data cleaning as our first work. For convenient, we remove nan column from original dataset. Next, we remove the duplicate data and none.

2) Feature selection: For feature selection, we have a lot of work to do. First, we remove all data whose "is_open" is 0. If some commerce is no longer open, there will be no sense to

consider such commerce. Besides, we remove the feature "business_id", which is used by Yelp but not helpful in our visualization. At last, since we only visualize state within U.S., we remove all data outside U.S. After feature selection, the features we remained for analysis are showed as below Chart 2.

feature name	description	data type
name	name of business	string
address	address of business	string
city	name of city	string
state	abbreviation of state	string
postal_code	postal code within U.S.	integer
latitude	latitude within U.S.	float
longitude	longitude within U.S.	float
stars	customer rating	float
review_count	count of reviews	integer
categories	1 category of business	string

Chart 2. Selected features for data analysis

C. Data analysis

Our programming is implemented by Python. To identify the "most popular within one state", we define the criterion as following:

- We select only one business in each state within U.S.
- Firstly, we sort the business by level of stars and then, sort them by count of customer reviews, both in decreasing. The highest one within each state is identified as "the most popular" one.
- If there are exactly some level of popularity within one state, we select randomly one.

After data analysis, we obtain a dataset as a result. In each state within U.S., there are only one business with information of its geography, location, customer rating and category.

IV. VISUALIZATION

In visualization part, we use HTML and CSS to build our webpage. Each state will have an icon on the top of the map standing for the best service within the state. When hovering over the icon, detailed information about the store, including store name, city and score will pop up, serving as a guide for people who'd love to travel in the state.

The screenshot of our visualization is showed as Figure1.

Figure 1. screenshot of visualization

V. CONCLUSION

There are still some challenges in our project. For example, Yelp dataset does not cover all states within U.S. Criterion of popularity is hard to determine. In visualization, information to show needs to be selected for efficiency. And there is some improvement could be done in the future work. For example, Search for more data which covers whole U.S. or even extend to global and combine stars and count of reviews by algorithms or machine learning model to obtain more objective result.

As for contribution of work, Zhuoyu was responsible for data collecting and processing part, which was obtaining store information from yelp, selecting data and locating them on google map. Yaofei did the visualization part on webpage. Each state will have an icon standing for the best local service on top of the map. When hovering over the icon, detailed information about this certain store, including name, city where it is located and score would pop up

REFERENCES

[1] https://www.kaggle.com/yelp-dataset/yelp-dataset/data