Group project report

Yulin Li
30526515
yl8n18@soton.ac.uk

Chen Tian
30072514
ct2n18@soton.ac.uk

Hang Su
00000000
hs1a18@soton.ac.uk

Zhihan Wang
00000000
zw3u18@soton.ac.uk

Jing Meng
00000000
jm4y17@soton.ac.uk

ABSTRACT

1 INTRODUCTION

2 CHOICES AND JUSTIFICATIONS 15POINTS

2.1 Reasoning behind the choice

Choosing a restaurant is one of the common problems faced by modern people on a daily basis. From the consumers point of view, making decision based on the reviews on restaurants rating applications may be a convenient choice. However, choosing those restaurants which have more reviews or higher stars could not satisfied us always. Can we trust current rating systems? How is the relationship between reviews and stars? Is there a way to provide a more reliable reference to customers before they make decisions? To answer these questions, we hope to find the relationship between customer reviews and the level of restaurants(one to five stars). We collect review data of all Canada restaurants from Yelp and optimize the rating algorithm by applying technologies and techniques we learn from lectures of foundation of data science.

2.2 Technologies and techniques

In this project, we use technologies including statistics, machine learning and natural language processing to build a restaurant recommendation system.

3 PROJECT PROCESS

3.1 Data collection and pre-processing

There are five datasets we used in the project are downloaded from Yelp. They range from business, checkin, review, tip to user data.

Firstly, we check each dataset seperately. The core business data file contains 188K business information including categories, city, review, stars and geographic information which are useful for our research. The review data file contains reviews data, stars data as well as user information, which can be combined with business data in the following steps. The check-in data, tip data and user data file contain more details including customer check-in time, tips for other users and average stars given by current users etc.

To achieve a relatively sufficient sample for the project, we decide to choose review data from cities of Canada which has more than 500 restaurants as our research object. After filtering city names based on geographic information (longitude and latitude) in business data file, we use business id to filter review data file. In a further step, we link business and review data files together and limit date to keep comments in 2017. Our cleaned data contains restaurants reviews from 9 cities in Canada with business id, name, categories and stars and other essential information. In addition, we keep user data file for the use of application visualisation.

3.2 Data exploration and analysis

In this step, we explore data in details. We are interested in the relationship between reviews and ratings of the restaurants, as a first step, we sort restaurants by the review count and list Top 50 restaurants with most reviews in Canada in 2017.[figure...] Furthermore, we take a look at the ratings of these 50 restaurants. We discovered that the average rate for all restaurants in Canada is 3.52, and the proportion of Top 50 restaurants with rate less than 4 and 3.5 is 0.24 and 0.12 respectively.[figure...]

After calculating the ratings, we use Pearson product-moment correlation coefficient to find out the whether the amount of reviews and average rating are related strongly or not. The result shows that the correlation coefficient is 0.15, which means that there is no significant relationship between the number of reviews and the ratings.[cite...]

Next, we want to explore the key features which have impact on the popularity of restaurants. We use tools in scikit-learn of python to extract Top 20 labels and their corresponding percentage as shown in Table 1 .

Label name	Percentage of all labels(%)
food	7.27
bars	6.29
nightlife	3.18
new	2.37
american	2.10
breakfast	2.07
brunch	2.07
japanese	2.02
chinese	1.78
canadian	1.69
cafes	1.38
tea	1.36
sandwiches	1.28
traditional	1.23
asian	1.22
italian	1.19
coffee	1.18
fusion	1.16
pizza	1.13
shshi	1.12

Table 1: Top 20 labels in the reviews

3.3 Rating and Recommend system building and evaluation

3.4 System visualization

4 ANALYSIS OF THE RESULTS 30POINTS

4.1 overall effectiveness

4.2 strength and weakness

5 TECHNICAL IMPLEMENTATION 35 POINTS

The application should be complete, should run without errors, and behave as defined.

6 REPORTING 20 POINTS

The report should be clear and professional in tone. High quality references should be used to justify statements.

6.1 —

7 CONCLUSIONS

REFERENCES