The Ohio State University, Department of Computer Science and Engineering

Homework 6: Prediction of rating from Yelp review text

Grover, Karan & Arora, Pragya & Ghai, Piyush

{grover.120, arora.170, ghai.8}@osu.edu

November 29, 2016

1 Problem Statement

1.1 About Yelp

Yelp [1] a famous website as well as a mobile app which publishes crowd sourced reviews about food joints and businesses. It also has a division which handles online reservations for restaurants. **Yelp Dataset Challenge**[2] is a publicly open contest sponsored by Yelp, in which the participants are challenged to use Yelp's data in an innovative way.

1.2 Our Mission

The Yelp dataset downloaded from Yelp dataset challenge website is huge and consists of over 2.7M reviews by roughly 687k users for over 86k businesses [2]. For this project, we chose to predict a review's rating based on the review text. The rating will be done on a scale of 1-5, where 1 stands for awful and 5 stands for excellent. We built multiple models and accessed which models would fit our use case the best. This is explained in more depth in the later sections of this report.

1.3 About the dataset

The Yelp Dataset consists of several files in JSON format of the data. The main files as per our use-case were the *yelp_academic_dataset_business.json* & *yelp_academic_dataset_review.json*. The data representation in the for acadmic_dataset_business is as follows :

```
{
    "type": "business",
    "business_id": (encrypted business id),
    "name": (business name),
    "neighborhoods": [(hood names)],
    "full_address": (localized address),
    "city": (city),
    "state": (state),
    "latitude": latitude,
    "longitude": longitude,
    "stars": (star rating, rounded to half-stars),
    "review_count": review count,
    "categories": [(localized category names)]
    "open": True / False (corresponds to closed, not business hours),
    "hours": {
        (day_of_week): {
            "open": (HH:MM),
            "close": (HH:MM)
        },
        ...
    },
    "attributes": {
        (attribute_name): (attribute_value),
    },
}
```

The data representation in the for acadmic_dataset_review is as follows :

```
{
```

```
        "type": "review",
    "business_id": (encrypted business id),
    "user_id": (encrypted user id),
    "stars": (star rating, rounded to half-stars),
    "text": (review text),
    "date": (date, formatted like "2012-03-14"),
    "votes": {(vote type): (count)},
}
```

The given datasets were first converted and exported into csv formats using a simple python script. The python script is a part of the **PreProcessing1.py** python file.

References

[1] ggplot Library is used in this assignment to plot most of the graphs in this assgnment `http://ggplot2.org/`.

[2] ggplot Library is used in this assignment to plot most of the graphs in this assgnment `http://ggplot2.org/`.