Customers’ Choice: An Yelp Case Study

[CSE 258 Assignment2]

Xianda Xie
A53218476
CSE, UCSD
xix100@ucsd.edu

Xiaowen Mao
A53220159
CSE, UCSD
x9mao@ucsd.edu

Pei Xu
A53204705
CSE, UCSD
pex007@ucsd.edu

ABSTRACT

Nowadays, when people decide where to go out for a dinner, especially when they are in a new place, they turn to Yelp for help. By reading other customers’ reviews and checking the stars of the restaurant, customers get to know the restaurant and decide whether to go. On the other hand, restaurants receive their credits each time customers rate on Yelp after dinning. What influence people’s preference and how should a restaurant make changes in order to attract more people? Those questions could be answered by analyzing the dataset on Yelp. By looking into the dataset, we find certain connections between restaurants’ features and customers’ ratings. In this paper, given the users’ reviews, we predicted the ratings of the restaurants over 4 countries based on a dataset derived from Yelp. We investigated classification models including Naive Bayes, linear regression, Latent-factor models and Logistic Regression to analyze their pros and cons on this prediction task.

Keywords

prediction;Linear regression; Latent-factor models; Logistic Regression

1. DATASET

We explored the dataset provided on Yelp website. This dataset contains detailed information regarding review, user and business respectively. We extract the first 600,000 samples to conduct our analysis.

1.1 Review Data Formula

The data of review is formulated as in Table 1.

It’s interesting that besides ”stars” attribute, there are also attributes like ”cool” and ”funny” that reveals other users’ attitude towards different aspects of this piece of review. So we performed an exploratory analysis of the data on the ”funny”,”cool”,”year” aspects of the review.

From the Fig 1-3 we can see that most reviews receive no vote in ”cool” and ”funny”. But those which has a higher

review_id	encrypted review id
user_id	encrypted user id
business_id	encrypted business id
stars	star rating, rounded to half-stars
date	date formatted like 2009-12-19
text	review text
useful	number of useful votes received
funny	number of funny votes received
cool	number of cool review votes received
type	review

Table 1: Review Data Formula

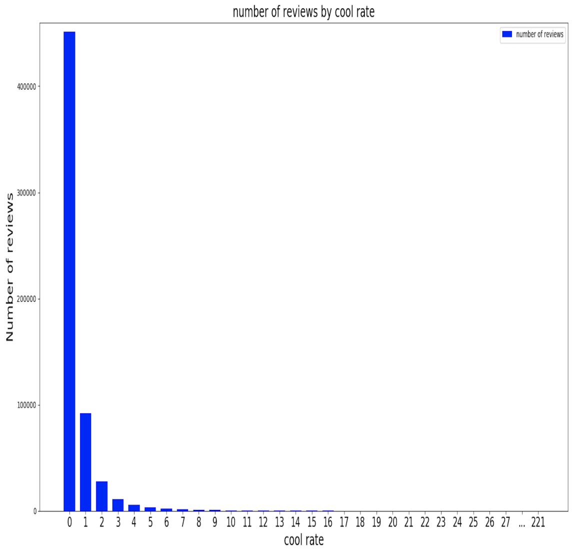

Figure 1: The relation between score in ”cool” and number of review

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

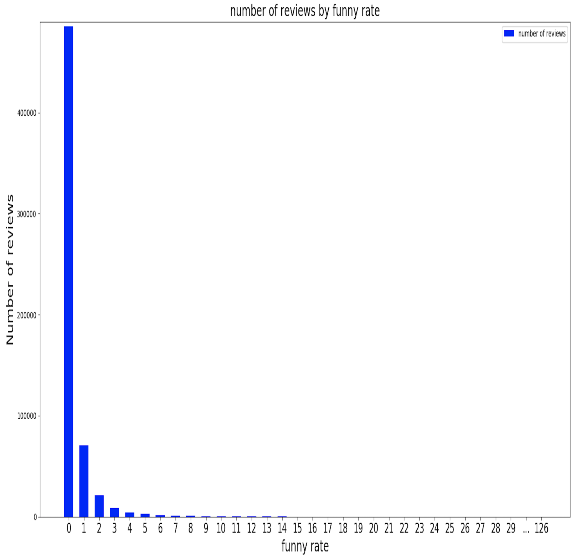

Figure 2: The relation between score in ”funny” and number of review

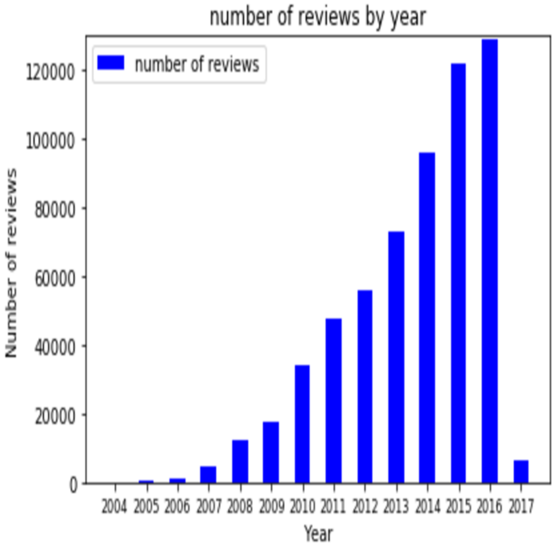

Figure 3: Different years have different numbers of reviews

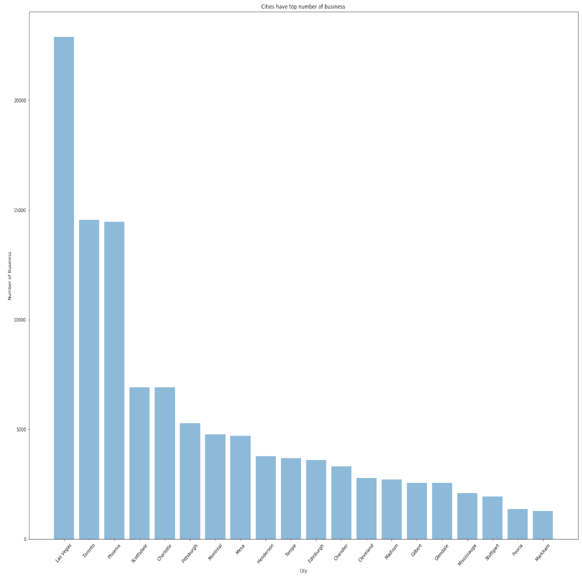

Figure 4: Cities with largest number of business

business_id	encrypted business id
name	business name
neighborhood	hood name
address	full address
city	city
state	state – if applicable –
postal code	postal code
latitude	latitude
longitude	longitude
stars	star rating, rounded to half-stars
review_count	number of reviews
is_open	0/1 (closed/open)
attributes	an array of strings: each array element is an attribute
categories	an array of strings of business categories
hours	an array of strings of business hours
type	business

Table 2: Business Data Formula

rating in stars tend to get a vote in ”cool” or ”funny”.

1.2 Business Data Formula

The data of business is formulated as in Table 2.

It’s obvious that ’stars’(star rating) is essential to predict star ratings of a specific review because users tend to rate a business around the average rating of this business.

Also, we find it interesting that the number of business of cities varies significantly. 20 cities with largest number of business can be seen in Figure 4. In addition, the number of business and the star ratings vary dramatically with different business categories. These data features, including the number of business, categories with highest average star rating and categories with lowest average star rating are shown in Figure 5, Figure 6 and Figure 7.

1.3 User Data Formula

The data of users is formulated as in Table 3.

user_id	encrypted user id
name	first name
review_count	number of reviews
yelping_since	date formatted like ”2009-12-19”
friends	an array of encrypted ids of friends
useful	number of useful votes sent by the user
funny	number of funny votes sent by the user
cool	number of cool votes sent by the user
fans	number of fans the user has
elite	an array of years the user was elite
average_stars	floating point average like 4.31
compliment_hot	number of hot compliments received by the user
compliment_more	number of more compliments received by the user
compliment_profile	number of profile compliments received by the user
compliment_cute	number of cute compliments received by the user
compliment_list	number of list compliments received by the user
compliment_note	number of note compliments received by the user
compliment_plain	number of plain compliments received by the user
compliment_cool	number of cool compliments received by the user
compliment_funny	number of funny compliments received by the user
compliment_writer	number of writer compliments received by the user
compliment_photos	number of photos compliments received by the user
type	user

Table 3: User Data Formula

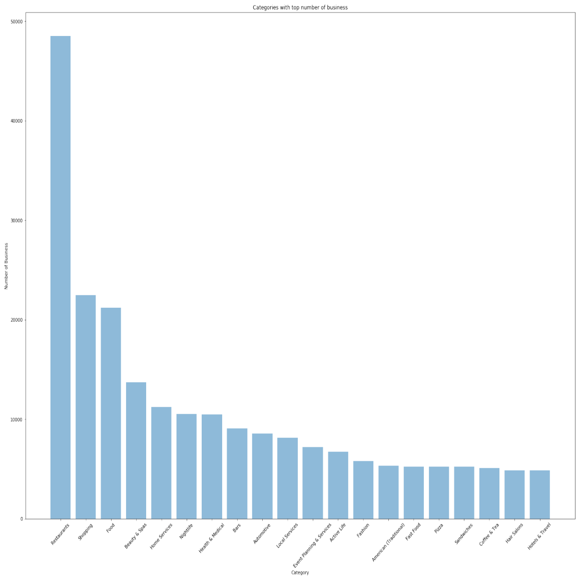

Figure 5: categories with largest number of business

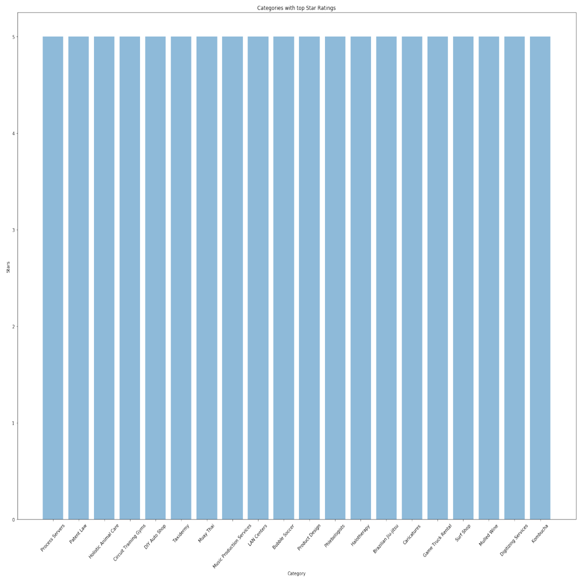

Figure 6: categories with highest average star ratings

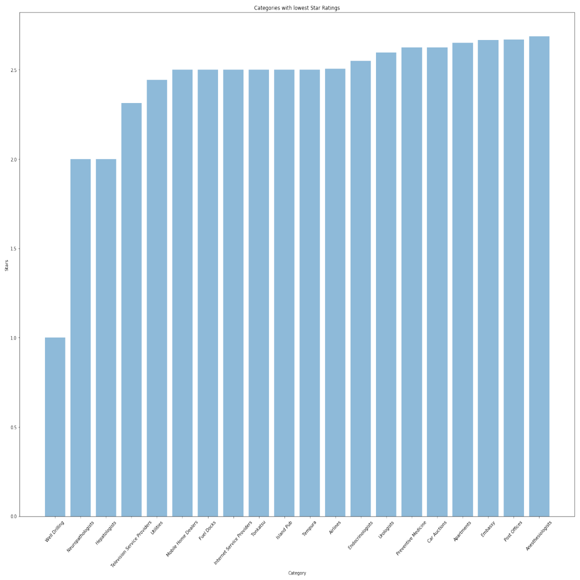

Figure 7: categories with lowest average star ratings

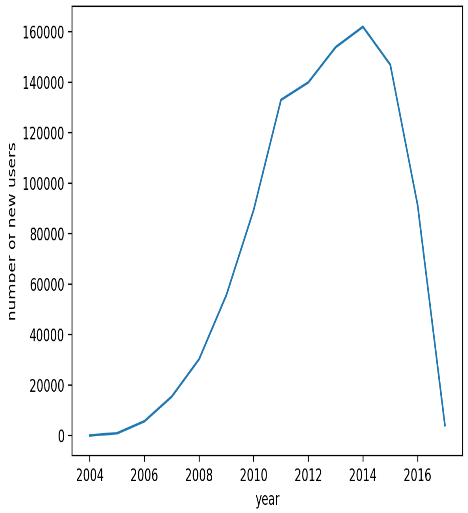

Figure 8: Different years with different number of newly registered users

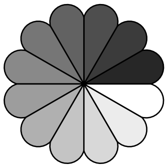

Figure 9: A sample black and white graphic that has been resized with the `includegraphics` command.

It's obvious that the average rating of a specific user has a great influence on the star rating he will give to business because users tend to give ratings around his average. In addition, the number of hot/plain/cool/funny compliments he has

1.3.1 Inline (In-text) Equations

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin. . .\end` construction or with the short form `$. . .$`. You can use any of the symbols and structures, from α to ω, available in LaTeX[8]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\rightarrow \infty}x=0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

1.4 Citations

Citations to articles [3, 5, 4, 6], conference proceedings [5] or books [9, 8] listed in the Bibliography section of your article will occur throughout the text of your article. You should use BibTeX to automatically produce this bibliography; you simply need to insert one of several citation commands with a key of the item cited in the proper location in the `.tex` file [8]. The key is a short reference you invent to uniquely identify each work; in this sample document, the key is the first author's surname and a word from the title. This identifying key is included with each item in the `.bib` file for your article.

1.5 Figures

Like tables, figures cannot be split across pages; the best placement for them is typically the top or the bottom of the page nearest their initial cite. To ensure this proper "floating" placement of figures, use the environment **figure** to enclose the figure and its caption.

This sample document contains examples of **.eps** files to be displayable with LaTeX. If you work with pdfLaTeX, use files in the **.pdf** format. Note that most modern TeX system will convert **.eps** to **.pdf** for you on the fly. More details on each of these is found in the *Author's Guide*.

As was the case with tables, you may want a figure that spans two columns. To do this, and still to ensure proper "floating" placement of tables, use the environment **figure*** to enclose the figure and its caption. and don't forget to end the environment with figure*, not figure!

1.6 Theorem-like Constructs

Other common constructs that may occur in your article are the forms for logical constructs like theorems, axioms, corollaries and proofs. There are two forms, one produced by the command `\newtheorem` and the other by the command

`\newdef`; perhaps the clearest and easiest way to distinguish them is to compare the two in the output of this sample document:

This uses the **theorem** environment, created by the `\newtheorem` command:

THEOREM 1. *Let f be continuous on $[a, b]$. If G is an antiderivative for f on $[a, b]$, then*

$$\int_a^b f(t)dt = G(b) - G(a).$$

The other uses the **definition** environment, created by the `\newdef` command:

Definition 1. If z is irrational, then by e^z we mean the unique number which has logarithm z:

$$\log e^z = z$$

Two lists of constructs that use one of these forms is given in the *Author's Guidelines*.

There is one other similar construct environment, which is already set up for you; i.e. you must *not* use a `\newdef` command to create it: the **proof** environment. Here is a example of its use:

PROOF. Suppose on the contrary there exists a real number L such that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = L.$$

Then

$$l = \lim_{x\to c} f(x) = \lim_{x\to c} \left[gx \cdot \frac{f(x)}{g(x)} \right] = \lim_{x\to c} g(x) \cdot \lim_{x\to c} \frac{f(x)}{g(x)} = 0 \cdot L = 0,$$

which contradicts our assumption that $l \neq 0$. □

Complete rules about using these environments and using the two different creation commands are in the *Author's Guide*; please consult it for more detailed instructions. If you need to use another construct, not listed therein, which you want to have the same formatting as the Theorem or the Definition[9] shown above, use the `\newtheorem` or the `\newdef` command, respectively, to create it.

2. PREDICTIVE TASK

2.1 Task Description

Identify a predictive task that can be studied on this dataset.

After our intensive data analysis that is done in the previous section, we would like to predict the star rating for a given review based on review features, user features and business features.

$$f(review\_feats, user\_feats, business\_feats) = stars$$

In the upcoming section we will discuss various prediction models with different feature selection.

2.2 Evaluation

We will evaluate your model at this predictive task using Mean Squared Error (MSE) for models:

$$MSE = \frac{1}{N} \sum_1^n y_i - f(feature_i)^2$$

The relevant baseline we used for comparison is the average star rating of the training set, rounded to half-stars. That is,

$$average = \frac{round(\frac{(\frac{1}{N}\sum_1^n stars_i)*10}{5}) * 5.0}{10}$$

The MSE of baseline is

$$MSE_{baseline} = 2.49221$$

To assess the validity of our model‘s predictions, we compare the MSE of baseline to the MSE of our model.

2.3 Models Selection

explain and justify which model was appropriate for the task.

2.4 Features Selection

ItâĂŹs also important in this section to carefully describe what features you will use and how you had to process the data to obtain them.

3. MODELS

Explain and justify your decision to use the model you proposed. How will you optimize it? Did you run into any issues due to scalability, overfitting, etc.? What other models did you consider for comparison? What were your unsuccessful attempts along the way? What are the strengths and weaknesses of the different models being compared?

3.1 Linear Regression

Regression is one of the simplest supervised learning approaches to learn relationships between input variables and output variables. Linear regression assumes a predictor of the form $X\theta = y$

3.2 Latent-factor Model

On the base of former models training with features, we went from a method which uses only features to one which completely ignores them, just like the professor mentioned in class.

1. Linear Model

$$star(business, user) = \alpha + \beta_b + \beta_u$$

2. Optimization

$$argmin_{\alpha,\beta} = \sum_{u,i} (\alpha + \beta_u + \beta_i - R_{u,i})^2 + \lambda_u \times \sum_u (\beta_u)^2 + \lambda_i \times \sum_i (\beta_i)^2$$

3. Iteration

$$\alpha^{(t+1)} = \frac{\sum_{u,i \in train} (R_{u,i} - (\beta_u^{(t)} + \beta_i^{(t)}))}{N_{train}}$$

$$\beta_u^{(t+1)} = \frac{\sum_{i \in I_u} (R_{u,i} - (\alpha^{(t+1)} + \beta_i^{(t)}))}{\lambda_u + |I_u|}$$

$$\beta_i^{(t+1)} = \frac{\sum_{u \in U_i} (R_{u,i} - (\alpha^{(t+1)} + \beta_u^{(t+1)}))}{\lambda_i + |U_i|}$$

4. Training

I trained my model using the whole 200,000 training dataset. After testing on Kaggle, I choose my parameters as following:

$$\lambda_u = 4.5$$

$$\lambda_i = 9.3$$

$$iterations = 30$$

And the validation MSE on 100,000 valid dataset is

$$MSE = 0.806253035$$

4. LITERATURE

In this section, we will describe some literature related to the problem we are studying.

4.1 About Dataset

We are using an existing dataset from Yelp Dataset Challenge Round 9, which can be downloaded from the website[11]. This set includes information about local businesses in 11 cities across 4 countries. The whole dataset is extremely large, containing 4.1M reviews and 947K tips by 1M users for 144K businesses, 1.1M business attributes (e.g., hours, parking availability, ambience), aggregated check-ins over time for each of the 125K businesses and 200,000 pictures from the included businesses. It is used to challenge students to use this dataset in an innovative way and break ground in research. There are several research topics can be done with this dataset, such as Cultural Trends, Location Mining and Urban Planning, Seasonal Trends, Infer Categories, Natural Language Processing(NLP), Changepoints and Events, Social Graph Mining, etc.

4.2 Other similar datasets

There are very few similar datasets for restaurants. Most of restaurant review datasets are composed of review messages only, with no information about funny or cool amounts to indicate their ratings. This kind of datasets are mainly used for Natural Language Processing(NLP) and text mining. For example, a Restaurant Review Dataset with 5531 restaurants and 52077 reviews[10], 2015 ABSA Restaurant Reviews with 254 reviews(1315 sentences)[1], etc.

And for other review dataset, the most popular one is Amazon Fine Foods reviews [2]. This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all 500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review, which is pretty similar with the Yelp one. The discussion and conclusions can be found here[7].

4.3 State-of-the-art Methods

Besides the research area of NLP mentioned before, there are other two different kind of methods currently employed to study this type of data. One is predicting sentiment, which makes up a prediction model to predict whether a recommendation is positive or negative. Other is word-cloud method, which is a popular method to represent boring large data as visualized graphs and animations.

4.4 Compared to ours

Compared to our own findings, the prediction from predicting sentiment model is more accurate. However, such kind of models can not figure out the most relevant result with ranking scores.

And we also tried the word cloud method on our review text. The result is really interesting, as Figure 4.

5. CONCLUSIONS—TODO

Describe your results and conclusions. How well does your model perform compared to alternatives, and what is the significance of the results? Which feature representations worked well and which do not? What is the interpretation of your modelâĂŹs parameters? Why did the proposed model succeed why others failed (or if it failed, why did it fail)?

6. REFERENCES

[1] ABSA. *2015 ABSA Restaurant Reviews.* http://metashare.ilsp.gr:8080/repository/browse/semeval-2015-absa-restaurant-reviews-train-data/b2ac9c0c198511e4a109842b2b6a04d751e6725f2ab847df88b19ea22cb5cc4a/.

[2] Amazon. *Amazon Fine Foods reviews.* https://snap.stanford.edu/data/web-FineFoods.html.

[3] M. Bowman, S. K. Debray, and L. L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.

[4] J. Braams. Babel, a multilingual style-option system for use with latex's standard document styles. *TUGboat*, 12(2):291–301, June 1991.

[5] M. Clark. Post congress tristesse. In *TeX90 Conference Proceedings*, pages 84–89. TeX Users Group, March 1991.

[6] M. Herlihy. A methodology for implementing highly concurrent data objects. *ACM Trans. Program. Lang. Syst.*, 15(5):745–770, November 1993.

[7] Kaggle. *Discussion on Kaggle.* https://www.kaggle.com/snap/amazon-fine-food-reviews/kernels.

[8] L. Lamport. *LaTeX User's Guide and Document Reference Manual.* Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.

[9] S. Salas and E. Hille. *Calculus: One and Several Variable.* John Wiley and Sons, New York, 1978.

[10] M. Sharifi. *Restaurant Reviews Dataset.* http://www.cs.cmu.edu/m̃ehrbod/RR/.

[11] Yelp. *Yelp Dataset Challenge.* https://www.yelp.com/dataset_challenge.

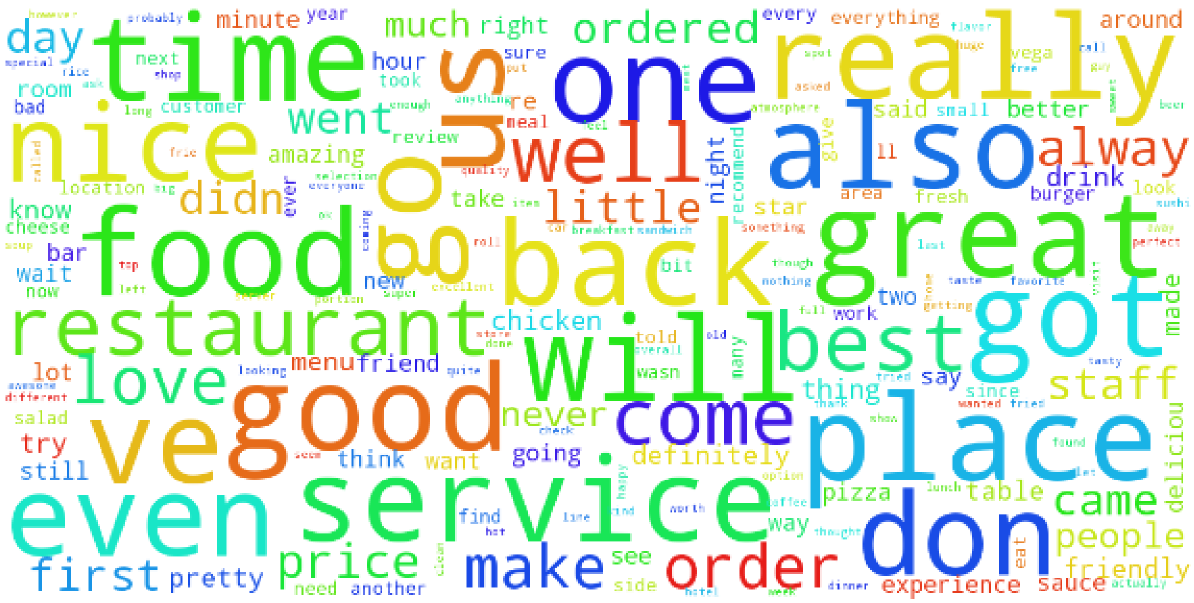

Figure 10: A sample word cloud graph on Yelp reviews.