Yelp Data Mining

Team members

Chris Powell

Jacob Bostick

Donald Boles

Description

- Dataset from Yelp, an online business rating website.
- Contains:
  - Business
  - Checkin
  - Photo
  - Review
  - Tip
  - User

Intended Questions

The Yelp dataset is a large and diverse set of information, which is why we chose to use it for the project.

Questions:

What types of reviewers are there? Can we identify ‘fake’ or ‘robot’ accounts?

What areas of the country that have the lowest ratings on Yelp?

Where should a person open a business based on these reviews in Yelp?

-Determine the type, location, name, and reviewers

How concentrated is the industry?

Prior Work

What types of reviewers are there? Can we identify ‘fake’ or ‘robot’ accounts?

Rahman, Mahmudur, et al. propose ways to exploit social, spatial and temporal signals in the yelp data set to detect fraudulent ratings. They were able to produce a 94% accuracy in classifying reviews as fake or not, and 95.8% accuracy in classifying misleading reviews. They do this through the use of Review Spike Detection (RSD) which is a technique that looks for spikes in positive or negative reviews then uses supervised learning to extract features from each review such as locality in reviews and friends linked to target fraudulent reviews.

- [1] Rahman, Mahmudur, et al. “To Catch a Fake: Curbing Deceptive Yelp Ratings and Venues.” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 8, no. 3, Oct. 2015, pp. 147–161., doi:10.1002/sam.11264. (https://users.cs.fiu.edu/~carbunar/deceptive.sam.pdf)

Banerjee, Shankhadeep, et al. found six main reviewer characteristics which supported their hypothesis (positivity, involvement, experience, reputation, competence, and sociability). They then uses those factors to manage a logistic regression model to classify review based on trustworthiness.

- [2] Banerjee, Shankhadeep, et al. “Whose Online Reviews to Trust? Understanding Reviewer Trustworthiness and Its Impact on Business.” *Decision Support Systems*, vol. 96, 2017, pp. 17–26., doi:10.1016/j.dss.2017.01.006.
  Whose Online Reviews to Trust link

Prior Work

What areas of the country that have the lowest ratings on Yelp?

Rahimi, Sohrab, et al. show that food selection, drink selection, and the general feel of restaurants can be good indicators of socioeconomic descriptors of neighborhoods

- [1] Rahimi, Sohrab, et al. “The Geography of Taste: Using Yelp to Study Urban Culture.” 2018, doi:10.20944/preprints201806.0389.v1.

Prior Work

Where should a person open a business based on these reviews in Yelp?

-Determine the type, location, name, and reviewers

M. Fan and M. Khademi demonstrates a method to predicts a business star from reviews. To determine the best forecast for choosing the restaurant category

- [1] M. Fan and M. Khademi, “Predicting a business star in yelp from its reviews text alone,” arXiv preprint arXiv:1401.0864, 2014.

T. Zhang and Y. Pan show a way to detect some of the effective facts about users, business and reviews, by using statistical analysis to determine the relationship between the success of the business and its geographic location.

- [2] T. Zhang and Y. Pan, “Yelp challenge project report,” 2014

W. O. Mengqi Yu, Meng Xue, predicted the rating for a restaurant from the user’s review histories and restaurant’s statistics

- [3] W. O. Mengqi Yu, Meng Xue, “Restaurants review star prediction for yelp dataset,” 2015.

Datasets

○ Within the Yelp dataset, we plan on using the reviewer, business, checkin, and user json files.

○ The Yelp dataset can be found at https://www.yelp.com/dataset/challenge.

○ Local download on all team members machines.

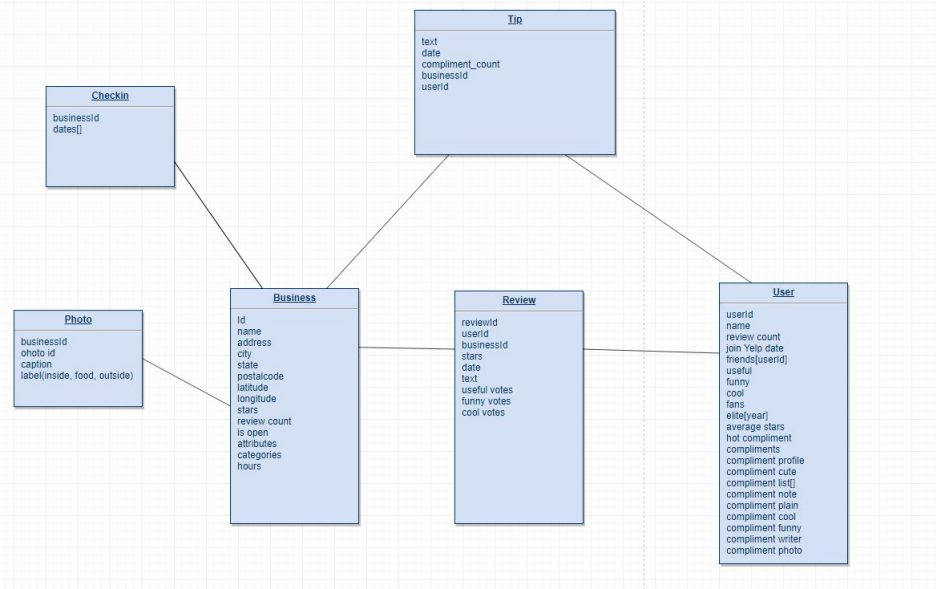

Diagram of json files and attributes

Proposed work

- Data cleaning: Removing null values using bins.
- Data preprocessing: Normalizing values to allow for comparisons.
- Data integration: Tie together the tables using the business Id and the ReviewID

List of tool(s)

- Programming Language: Python
- Libraries: pandas, plot.ly
- Repository: github

Evaluation

1. Unsupervised learning on clustering users to detect robot or fake accounts.
2. Normalizing data for use in classification methods.
3. Pattern mining for () -> positive review, for the attributes of name, type, description.
4. Use bin discretization for categorical variables to be converted for use in classification and prediction.
5. Use outliers and clustering to remove data points from consideration.
6. Use contextual outliers to remove outliers when considering locations as average reviews.
7. Use K-nearest neighbor on our guess for a new business to see how it appears in a visualization with the attributes together after we mine the individual traits.