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?

[1] https://users.cs.fiu.edu/~carbunar/deceptive.sam.pdf

[2] Whose Online Reviews to Trust link

Prior Work

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

[1] file:///Users/2015mbp16gb256gb/Downloads/The_Geography_of_Taste_Using_.pdf
[2] Your Neighbors Affect Your Ratings Link

Prior Work

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

-Determine the type, location, name, and reviewers

[1] https://ieeexplore-ieee-org.colorado.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=8126196
https://www.kaggle.com/ksjpswaroop/yelp-data-analysis: help existing business owners, future business owners to make important decisions regarding new business or business expansion

https://www.kaggle.com/ypaudel/extensive-eda-on-yelp-business-data

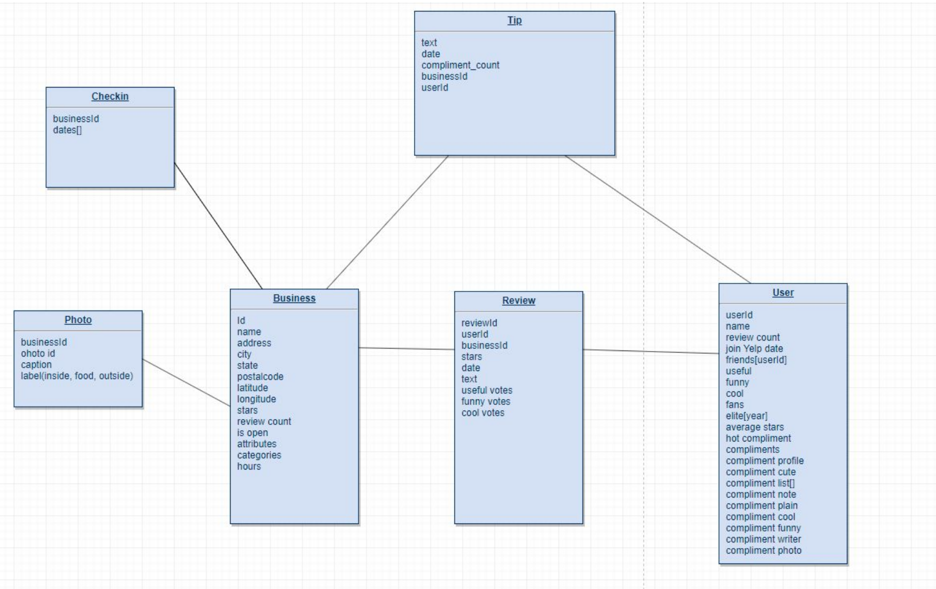
M. Fan and M. Khademi, “Predicting a business star in yelp from its reviews text alone,” arXiv preprint arXiv:1401.0864, 2014.
T. Zhang and Y. Pan, “Yelp challenge project report,” 2014
Q. Jin, “A research proposal: The effects of restaurant environment on consumer behavior,” in MBA Student Scholarship. 36., 2015.
W. O. Mengqi Yu, Meng Xue, “Restaurants review star prediction for yelp dataset,” 2015.
E. J. James Huang, Stephanie Rogers, “Improving restaurants by extracting subtopics from yelp reviews,” 2013.

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.



Tip
text
date
compliment_count
businessId
userId

Checkin
businessId
dates[]

Photo
businessId
ohoto id
caption
label(inside, food, outside)

Business
Id
name
address
city
state
postalcode
latitude
longitude
stars
review count
is open
attributes
categories
hours

Review
reviewId
userId
businessId
stars
date
text
useful votes
funny votes
cool votes

User
userId
name
review count
join Yelp date
friends[userId]
useful
funny
cool
fans
elite[year]
average stars
hot compliment
compliments
compliment profile
compliment cute
compliment list[]
compliment note
compliment plain
compliment cool
compliment funny
compliment writer
compliment photo

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.