CSE6242 Data and Visual Analytics

Project: Yelp Data Visualization for Analytics

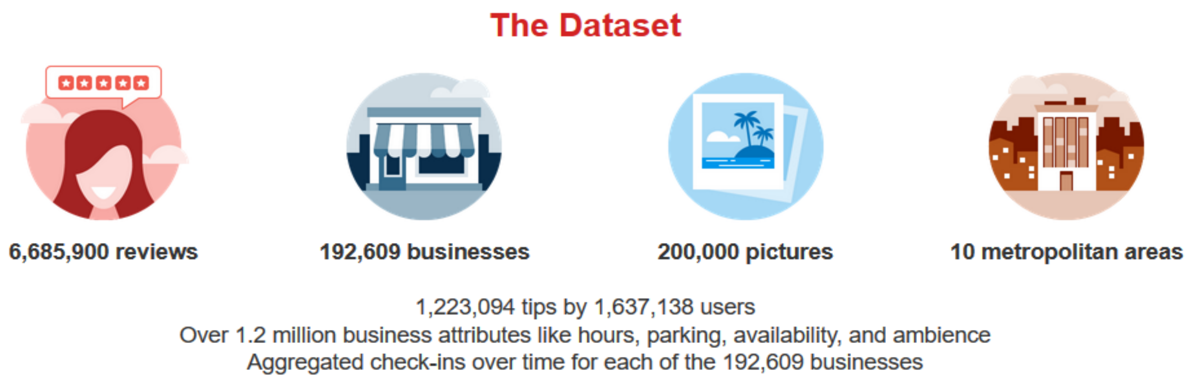
Team 145:
Karthik Balasubramanian (kbalasub6), Md Khaled Hassan (mhassan49), Suneil G Shrivastav (sshrivastav07), Ashutosh Verma(averm317), and Tianji Zhou (tzhou91)

I. Introduction

The aim of this project is to use very large dataset from Yelp(~8.6GB) and provide meaningful visualizations/dashboard to business analysts, restaurant owners & economic policy makers. The dataset contains multitude of indicators like user reviews, ratings, co-ordinates, zip code, review counts, categories etc. which we intend to correlate and produce such visualizations. We also intent to use the poverty/ population dataset per county provided in homework-2(HW2) for such visualizations.
Some examples of visualizations but not limited to

- Poverty data vs number of restaurants vs ratings per county
- Average restaurant rating per county
- Mapping restaurant & review frequency by geography
- Food diversity/popular attributes/ available facilitates
- Text analysis of reviews to provide actual insights
- Attempt to differentiate between genuine and fake reviews
- Restaurant rating over a period of time and factors impacting the rating
- Type of cuisines (ethnic / healthy / chains / …) in each geographic location

Note: Our analysis will be affected by the availability of data points in the original Yelp dataset



The Dataset

6,685,900 reviews

192,609 businesses

200,000 pictures

10 metropolitan areas

1,223,094 tips by 1,637,138 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 192,609 businesses

II. **Literature Survey:**

Review 1 & 2: Reviews, Reputation, and Revenue: The Case of Yelp.com [1],
Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud [2]

Main Idea: In [1], the authors combined Yelp.com dataset and restaurant data from the Washington State Department of Revenue to evaluate the correlation between yelp rating and restaurants revenue and how that can dominate the exit behaviors of restaurants. the key finding from this work is one yelp start increase can lead to 5-9% increase in revenue. In [2], the authors carried out similar empirical analysis and concluded that nearly 1 out 5 reviews are marked as fake by Yelp and restaurants with weak reputation are more likely to commit review fraud .
Why/why not useful for us: The strong correlation between star rating and revenue increase, and the volume of review frauds committed by some of the restaurants as shown in these papers provide us with the motivation for this work since revenue is a key determinant of a restaurant's decision to exit and the fake reviews can significantly impact that.
Potential shortcomings: none

Review 3: Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews [3]

Main Idea: In order to find a good restaurant rather than relying on the overall rating from Yelp dataset, word scores(frequency) generated from SVM(support Vector Machine) model to extract sentiment tendency of each review will yield better results. These word scores are bagged into positive and negative sentiments for visualization.
Why/why not useful for us: Useful to provide actual insights to restaurant owners.
Potential shortcomings: Focussed only on text analysis.

Review 4:Improving Restaurants by Extracting Subtopics from Yelp Reviews [4]

Main Idea: In this paper authors use Latent Dirichlet Allocation (LDA) algorithm (unsupervised) to discover hidden subtopics in user reviews to provide meaningful insights to restaurant owners about what customer cares about impacting their Yelp ratings. It turns out that Service,value,takeout and decor were the key indicators.
Why/why not useful for us: Useful as the key indicators from this paper will be used in charting the data.
Potential shortcomings: None.

Review 5: The Geography of Taste: Using Yelp to Study Urban Culture [5]

Main Idea: In this paper authors use Bourdieu's theory of distinction which emphasize food, drink, & interior decoration as best indicators of taste reflecting one's everyday choice. These indicators are extracted from yelp reviews using NLP(Natural Language Processing)techniques and statistically tied to demographic factors like income, racial composition and education to reflect socio-economic status of population in different neighbourhoods .
Why/why not useful for us : Useful as one aspect of our visualization is poverty data vs number of restaurant vs ratings
Potential shortcomings: limited focus as based only on specific indicators.

Review 6: Spatial analysis of users-generated ratings of yelp venues [6]

Main Idea:
The paper explored geographic pattern based on Yelp user ratings for different venues such as restaurants, fast foods, bars, etc. The authors filtered out and ignored venues that have less than 10 rating counts. The paper identified hot and cold spots (based on cluster value or average star rating) using AMOEBA algorithm. The study concludes that the rating based mapping gives insight on spatial patterns of service quality of business venues.
Why/why not useful for us:
Can leverage this approach to both star and text based rating.
Potential shortcomings:
Assumed venues with less than 10 ratings are likely to have fraud reviews.
Focused on only one city.
Need to consider both star and text ratings to detect fraud reviews.

Review 7: Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity [7]

Main Idea: Assess yelp's predictive power (using random forest algorithm) to predict the growth in CBP (County Business Patterns) establishments. The authors concluded that yelp provides more accurate prediction of local economy when population density is high.
Why/why not useful for us: we can leverage Yelp's predictive power to find correlation between poverty level (by county) and business models.
Potential shortcomings: Yelp's prediction power is limited in low density areas (due to low number of ratings available).

Review 8: Restaurant Setup Business Analysis Using Yelp Dataset [8]

Main Idea: The authors proposed a framework for restaurant business analysis. Using the framework, they first identified the most desired attribute in which a customer is most interested. In addition, they identified the most crowded day of the week and finally using kd-tree algorithm, they identified the nearest 10 restaurants for a given latitude and longitude to understand how the location can affect the business and to provide the most desired services, facilities, and cuisines.
Why/why not useful for us: We can provide a list of desired attributes and facilities to set up restaurant for each US county.
Potential shortcomings: Analysis is limited to a given search coordinates. In addition, most inactive day of the week needs to be identified to keep the restaurant closed.

Review 9: Restaurants Review Star Prediction for Yelp Dataset [9]

Main Idea: The authors attempted to predict restaurant ratings based on the user's review histories and restaurant's statistics using various machine learning techniques combining with the sentiment analysis. The authors concluded that random forest algorithm gives most accurate prediction.
Why/why not useful for us: Similar machine learning algorithms and sentiment analysis can be applied to our work.
Potential shortcomings: Analysis is limited to start prediction and not useful for future restaurant owners.

Review 10-16 (book chapters):
To explore possible machine learning and text based analysis we refer to a classic book in this field, ***Learning From Data – A Short Course*** [10]. **Chapter 1** of this book defines the learning problem. Different learning problems are grouped into three types: supervised learning, unsupervised learning and reinforcement learning. As our yelp dataset contains rating information, we will be mainly working with supervised learning, although we will also extract patterns by clustering and other unsupervised learning algorithms. **Chapter 2** provides a theoretical background of learning, by looking at the mathematical analysis of generalization, especially VC dimension of a hypothesis set. Intuitively, the more hypothesis one has, the easier it is to fit the training data, but overfit as well. The approximation-generalization trade-off must be considered, and we plan to break our dataset into training data and testing data, before showing our findings to our audience. **Chapter 3** introduce machine learning practice with most simple linear models such as linear classification/perceptron, linear regression and logistic regression. These algorithms are foundations of more complicated algorithms, and can

be generalized to non-linear features by nonlinear transformation. For our Yelp data we will consider linear regression as our label/output would be the rating, which is a continuous variable. **Chapter 4** dives into the topic of overfitting a bit more in depth. As the dataset always contains error and noise, which is especially true for our big Yelp dataset, trusting the training data too much will lead to overfitting and compromised prediction power on testing data. In practice this can be avoided by cross validation, which is to use part of the training data as validation set. In our case we will solve this issue by constructing our hypothesis space with reasonable number of hypothesis, such that given the size of our dataset we can eliminate overfitting. **Chapter 5** describes three machine learning principles: Occam's razor, sampling bias, and data snooping. In short, one should introduce features only when they are necessary, otherwise the hypothesis set will be so large that one overfits; if data is sampled in a biased way the learning will be also biased; looking at the data before choosing hypothesis is going to compromise learning. Our taken away from this Chapter is to analyze and understand the attributes and select them first before applying any hypothesis/learning. The rest of the chapters in this book describes more advanced machine learning techniques. **Chapter 6** introduces clustering algorithms such as k-nearest neighbors(kNN) and radial basis functions(RBF). We will apply kNN clustering to our dataset. **Chapter 7** introduces Neural Networks, which is intuitively multilayer perceptron, but can also be developed into deep neural networks such as convolutional neural networks. For our data set we think we will apply linear algorithms and also neural networks to do regression. Chapter 8 describes support vector machines, and Chapter 9 provide further reading and learning aids.

III. Nine Heilmeier questions:

1. What are you trying to do? Articulate your objectives using absolutely no jargon.

The aim of this project is to use very large dataset from Yelp(~8.6GB) and provide meaningful visualizations/dashboard to Business analysts, restaurant owners & economic policy makers. The dataset contains multitude of indicators like user reviews, ratings, co-ordinates, zip-code, review counts, categories etc. which we intend to correlate and produce such visualizations. We also intent to use the poverty/ population dataset per county provided in HW2 for such visualizations.More details in introduction.

2. How is it done today; what are the limits of current practice?

Similar studies have been carried on with limited number of attributes in this dataset, and we intend to provide a more comprehensive analysis with additional geographical and demographic data.

3. **What's new in your approach? Why will it be successful?**

Text based analysis, sentiment analysis, correlation with HW2 poverty dataset in addition to comprehensive list of charts via geographic and demographic data for through business analysis is the novelty of this project and hence will be successful.

4. **Who cares?**

Our study will provide detailed insights to restaurant owners, customers, suppliers, economists and business analysts.

5. **If you're successful, what difference and impact will it make, and how do you measure them (e.g., via user studies, experiments, ground truth data, etc.)?**

From the papers we understand that an increment of 1 star in the user rating brings in significant revenue. If the predictive analysis is successful it will directly impact business growth .

6. **What are the risks and payoffs?**

One risk is our analysis is dependent on Yelp data credibility. Another risk is, such analysis could make business investments worse in economically challenging areas.

The analysis will directly impact business investments, trigger corrective actions and hence revenue growth.

7. **How much will it cost?**

This is an educational project and hence free.

8. **How long will it take?**

Approximately 5 weeks.

9. **What are the midterm and final "exams" to check for success? How will progress be measured.**

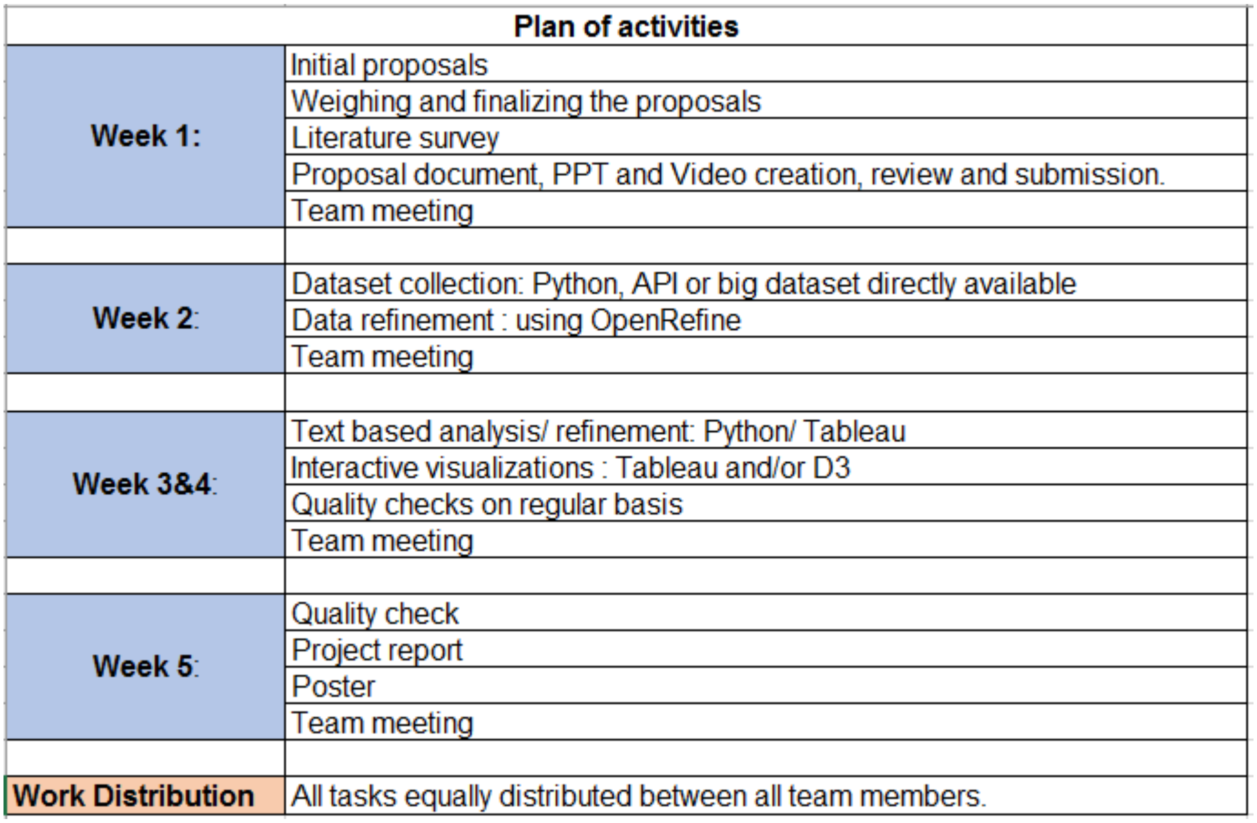
Midterm:

- Data collection and refinement
- data connected to county data; text based refinement
- a few straightforward visualizations

Final:

Working visualizations and other deliverables; project report; poster.

IV. Plan of activities



Plan of activities	
Week 1:	Initial proposals
	Weighing and finalizing the proposals
	Literature survey
	Proposal document, PPT and Video creation, review and submission.
	Team meeting
Week 2:	Dataset collection: Python, API or big dataset directly available
	Data refinement : using OpenRefine
	Team meeting
Week 3&4:	Text based analysis/ refinement: Python/ Tableau
	Interactive visualizations : Tableau and/or D3
	Quality checks on regular basis
	Team meeting
Week 5:	Quality check
	Project report
	Poster
	Team meeting
Work Distribution	All tasks equally distributed between all team members.

References:

[1]. *Reviews, Reputation, and Revenue: The Case of Yelp.com,* L. Michael, Harvard Business School NOM Unit Working Paper No. 12-016 (2016).

[2] *Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud*, Luca, Michael & Zervas, Georgios. (2013). SSRN Electronic Journal. 10.2139/ssrn.2293164.

[3]. *Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews*, Yu B., Zhou J., Zhang Y., Cao Y., arXiv e-prints, arXiv:1709.08698 (2017).

[4]. *Improving Restaurants by Extracting Subtopics from Yelp Reviews*, Huang, J.; Rogers, S.; Joo, E. Improving Restaurants by Extracting Subtopics from Yelp Reviews. In iConference Social Media Expo (2014).

[5]. *The Geography of Taste: Using Yelp to Study Urban Culture*, Rahimi, S.; Mottahedi, S.; Liu, X. *ISPRS Int. J. Geo-Inf. 7*, 376(2018).

[6]. *Spatial analysis of users-generated ratings of yelp venues*, Sun, Y. & Paule, J.D.G. Open geospatial data, softw. stand. https://doi.org/10.1186/s40965-017-0020-9 (2017).

[7]. *Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity* E. L. Glaeser, H. Kim, M. Luca, National Bureau of Economic Research,Working Paper No. 24010 (2017).

[8]. *Restaurant Setup Business Analysis Using Yelp Dataset* , Sindhu B Hegde, Supriya Satyappanavar, Shankar Setty, Advances in Computing Communications and Informatics (ICACCI) 2018 International Conference on, pp. 1455-1462, (2018).

[9]. *Restaurants Review Star Prediction for Yelp Dataset*, M. Yu, M. Xue, W. Ouyang, pdfs.semanticscholar.org (2015).

[10]. *Learning from data*. **Chapters 1-9** Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. New York, NY, USA:: AMLBook, 2012.