Machine Learning Final Project:

Exploring Yelp Reviews Using Topic Modeling, Sentiment Analysis and Recommender Systems

Alex Teboul, Brooks Thorton, JiaHao Chuah

DePaul University

Professor R.B. Tchoua

DSC 478

November 26, 2019

Table of Contents

1. Executive Summary

- **1.1. Project Goal……**
- **1.2. Methodology……....**
- **1.3. Conclusion……...**

2. Introduction

- **2.1. Dataset……**
- **2.2. Proposal……...**

3. Data Preprocessing and Exploration

- **3.1. Data Cleaning & Transformations……...……**
- **3.2. Data Exploration……**
- **3.3. Methods……....**

4. Methods & Results

- **4.1. Topic Modeling - LDA and LSA Analysis…….....**
  JiaHao……...
- **4.2. Sentiment Analysis & Model Building…….…..**
  Alex……....
- **4.3. Recommender Systems……...**
  Brooks……...

5. Discussion & Future Work

- **5.1. Discussion……...……**
- **5.2. Future Work……...……**

6. Appendices

1. Executive Summary

1.1. Project Goal

Yelp is a business directory source and a crowd-sourced review platform for businesses. There are a massive number of reviews for a multitude of business categories from all around the world on the Yelp platform. Looking through text reviews to gain business insights and feedback is a daunting and time-consuming task. It is important for Yelp and business on the platform to be able to mine data from business reviews to infer meaning, business attributes, and sentiment. More broadly, being able to understand the meaning of a body of text computationally has a number of useful applications. Our analysis of the Yelp dataset seeks to address this challenge of machine understanding of text data.

Our team aims to use various Natural Language Processing (NLP) techniques and Recommender System algorithms on Yelp's review dataset in order to better understand the content of reviews, their sentiment, relationship with ratings, and how best to provide future recommendations.

1.2. Methodology

In our analysis, we focus on a subset of the Yelp dataset. Specifically, we created a subset comprised of reviews from performing arts businesses in Las Vegas, Nevada.

We focused on three tasks:

1. **JiaHao.** Conduct topic modeling and feature selection to extract important topics from reviews using Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA)
2. **Alex.** Perform Sentiment Analysis to explore how text in reviews is related to ratings.
3. **Brooks.** Design a Recommender System through Collaborative filtering to suggest future businesses based on past user ratings.

For the topic modeling and sentiment analysis portion we first performed Exploratory Data Analysis to understand our dataset, and combine the dataset (business, review, user) based on the performing arts businesses in Las Vegas. We renamed the attributes, checked for missing values, and dropped attributes that did not provide useful information for our analysis. Our final dataset was comprised of 37,648 observations and 12 columns. We named this dataset vegas_cleaned.csv.

The recommender system portion of the analysis utilized data specific to Las Vegas as well, while also including all businesses across each category in addition to the performing arts category. Only users and businesses with a minimum of 10 reviews were utilized to build the recommendation system. The final data included contained 7,825 total reviews, by 251 users, for 278 businesses. This data set was named reviews_subset.csv.

JiaHao. We then performed Topic modeling analysis using LDA and LSA. After cleaning and preprocessing the data to fit to a doc-term matrix, we used LDA to find repeating terms and

extracted the top 50 important topics, and visualize the important terms and their frequencies across each topic using the pyLDAvis library. Then, we used LSA to leverage the context around the words in the reviews to capture the hidden concepts or topics. We created the doc-term matrix using TfidfVectorizer from the sklearn library, and represented the matrix as a vector by using truncatedSVD to print out the top 10 important words for each topic. We also performed LSA on low rating reviews (rating < 4) vs. high rating reviews (rating >= 4). We compared results from the LDA and LSA analysis to gain insights on important words from each topic, look at the positive and negative terms that pertains to the performing arts businesses in Vegas, and evaluate them with the actual reviews.

Alex. Following LDA and LSA, we decided to explore Yelp Reviews using Sentiment Analysis. This method is broken out into two parts. In Part 1, we wanted to explore how the body of text in a yelp review and yelp ratings are related. Specifically looking to determine if sentiment from text could be extracted and compared to ratings. We investigated how well an existing sentiment analysis package, TextBlob, could work to analyze sentiment in yelp reviews. We also created word clouds of the most frequently used terms for high versus low ratings. In Part 2, we explored sentiment model building using the IMDB movie review and applied the model to our Yelp reviews dataset. We wanted to determine if a sentiment model built on an entirely different dataset would still work effectively on our Yelp dataset. To accomplish this, we trained a multilayer sequential model using the Keras package on the accompanying IMDB movie review dataset. This model was then applied to text from Yelp reviews. Finally, we tested how well sentiment scores could predict whether a review received a high rating (4 or 5 stars) vs low rating (1, 2, or 3 stars) using Logistic Regression. We found that sentiment was highly predictive of rating, offering the potential for dimensionality reduction and prediction based on review text.

Brooks. Finally, we looked into how the Yelp review data set could be used to make future recommendations to each of the users in the Las Vegas area. Of primary focus was how the Collaborative Filtering based approach to recommendations could be utilized to build powerful recommendations. Singular Value Decomposition (SVD) was originally utilized to reduce the feature space and a total of 50 components were found through the utilization of the technique. Various recommender algorithms were employed which were measured in terms of accuracy and computational efficiency. The highest performing algorithm included SVDpp, SVD, KNNBaseline, and KNNWithMeans. The SVDpp algorithm was further analyzed and enhanced via parameter turning through the use of grid-based searching. The process produced a highly tuned SVDpp parameter adjusted algorithm which was tested on several user use cases.

1.3. Conclusion

JiaHao. The LDA/LSA analysis shows that we can identify the important patterns on positive and negative terms from the topics, which corresponds to the actual reviews. It shows that LDA/LSA allows us to gain useful insights on the important topics without going through the trouble reading reviews individually. We conclude that users are generally happy with the pinball machine games, fountain shows, Cirque du Soleil shows, music, and the businesses' effect on family entertainment, but are generally unsatisfactory with wasting money on boring, disappointing shows.

Alex. From Part 1 of the Sentiment Analysis we learned that we can determine a sentiment score for Yelp reviews using the TextBlob package. This sentiment score was moderately correlated with review_stars and with our highlow rating binary variable. We also found that there were discernible differences in the most frequently used words for reviews with high versus low ratings. Specifically, our WordClouds show words like 'canceled', 'worst', and 'wait' in the low ratings group and words like 'wonderful', 'absolutely', and 'best' in the high ratings group. This confirmed that we were on the right track with our analysis. In Part 2, we determined that a sentiment model trained using the IMDB dataset could be used successfully to give Yelp Reviews a sentiment score. This score between 0 (negative sentiment) and 1 (positive sentiment) could then be used to predict whether a review was rated highly or lowly with 83.8% accuracy, using our logistic regression model. This supports our theory that sentiment analysis performed on a large enough dataset of reviews can be extrapolated to other datasets and fields with a reasonable high degree of accuracy.

Brooks. The results of the Recommender System through Collaborative filtering demonstrated how different recommender algorithm can be utilized on sparse data sets and produce accurate recommendations for users. The Las Vegas reviews data was found to be ideal for the Collaborative Filtering based approach. In terms of accuracy the SVDpp algorithm with parameter adjustments produces the best performing recommender system for the Yelp users.

2. Introduction

2.1. Dataset

Yelp is a business directory source and a crowd-sourced review platform for businesses. There are a massive number of reviews for a multitude of business categories from all around the world on the Yelp platform. Looking through text reviews to gain business insights and feedback is a daunting and time-consuming task. It is important for Yelp and business on the platform to be able to mine data from business reviews to infer meaning, business attributes, and sentiment.

Our team aims to use various Natural Language Processing (NLP) techniques and Recommender System algorithms on Yelp's review dataset in order to better understand the content of reviews, their sentiment, relationship with ratings, and how best to provide future recommendations.

Our dataset was obtained from the yelp dataset challenge website. The challenge allows students to conduct research/analysis and share discoveries. The challenge provides us various dataset in .json formatted files. We used only the dataset that are related to the reviews, which are namely:

1. Business.json – contains business data including location data, attributes, and categories
2. Review.json – contains full review text data including the user_id that wrote the review and the business_id the review is written for.
3. User.json – user data including the user's friend mapping and all the metadata associated with the user.

Dataset Link: https://www.yelp.com/dataset/challenge

Since this would be our initial NLP analysis, we decided to target reviews from Las Vegas, Nevada's performing arts businesses, since there are a variety of entertainment in Vegas in which we can garner different insights, and allows us to work with a sizeable dataset.

2.2. Proposal

We propose to use various Natural Language Processing (NLP) techniques and Recommender System algorithms on Yelp's review dataset in order to better understand the content of reviews, their sentiment, relationship with ratings, and how best to provide future recommendations.

3. Data Preprocessing and Exploration

3.1. Data Cleaning & Transformations

EDA was performed on the dataset based on Yelp reviews for different businesses. The business dataset was first looked at to identify the locations that we would be conducting analysis from. Since the dataset provided has no businesses listed in Chicago, we decided to work with businesses in Las Vegas, since it has a wider variety of business categories, and would provide us a more diverse reviews. To control the size of our dataset, we tested out a few business categories, such as restaurants, nightlife, entertainment etc. We eventually settled on performing arts. The category provides us a sizeable number of business (436 observations), in which we combine the dataset with the review dataset and the user dataset.

Once the dataset was combined, we have to rename the attributes and check for missing values. We then dropped attributes that does not provide useful data for our analysis, such as useful, funny, cool reviews, business hours, user yelp history, and votes. Our final dataset comes out to 37648 observations and 12 columns. Our attributes are as follows:

Table 1: Dataset Description and Attributes

Attributes	**Description**	**Example**
text	string: review text	'Like walking back in time, every Saturday morning my sister and I was in a bowling league and after we were done, we'd spend a few quarters playing the pin ball machines until our mother came to pick us up.
date	date/time: date and time when the review is posted	11/30/2011 2:11:15 AM
review_stars	float: review star rating	4
business_name	string: business name	Pinball Hall Of Fame
address	string: business address	1610 E Tropicana Ave

business_stars	float: business average star rating	4.5
is_open	boolean: whether the business is still operating	1
attributes	boolean/dict: business attributes	{'RestaurantsGoodForGroups': 'True', 'RestaurantsDelivery': 'False', 'BusinessAcceptsCreditCards': 'False', 'OutdoorSeating': 'False', 'Alcohol': "u'none'", 'BusinessParking': "{'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': False}", 'RestaurantsReservations': 'True', 'RestaurantsAttire': "u'casual'", 'RestaurantsPriceRange2': '1', 'RestaurantsTakeOut': 'False', 'BikeParking': 'True', 'GoodForKids': 'True'}
categories	list: business categories	Performing Arts, Amusement Parks, Museums, Arcades, Arts & Entertainment, Active Life, Restaurants
user_name	string: user name for the review	Carol
user_review_count	integer: number of reviews that the user has given	866
average_stars	float: average star ratings for the user	4.16

Different attributes are used for the different analysis. Since we are focused on NLP analysis, we prioritize attributes text and review_stars. We will also be conducting analysis on reviews with low (< 4) and high (>= 4) ratings. However, we hope to gain insight from other attributes based on the results we find in our analysis as well.

3.2. Data Exploration

Various steps were taken exploring the dataset, some of which were mentioned above. Figure 1, presented below details some descriptive statistics on our numeric variables. Note that there is a skew in the ratings towards more positive scores, with an average rating of 4.0. We will be working with the 37,648 reviews for the methods that follow.

	review_stars	business_stars	is_open	user_review_count	average_stars
count	37648.000000	37648.000000	37648.000000	37648.000000	37648.000000
mean	4.023720	4.028806	0.872875	208.732974	3.790983
std	1.309937	0.650636	0.333117	443.668726	0.634756
min	1.000000	1.500000	0.000000	1.000000	1.000000
25%	3.000000	3.500000	1.000000	15.000000	3.500000
50%	5.000000	4.000000	1.000000	56.000000	3.820000
75%	5.000000	4.500000	1.000000	211.000000	4.150000
max	5.000000	5.000000	1.000000	12390.000000	5.000000

Figure 1: Performing Arts Yelp Reviews for Vegas - Descriptive Statistics

As part of the exploratory process in the Sentiment Analysis portion of our methods, we explored word clouds formed from the text of high versus low ratings. Figure 2 shows the word cloud for low ratings while Figure 3 show it for high ratings. Note that some words make sense in each category, with words like 'worst' in the low ratings and 'wonderful' in the high ratings. There are others that appear misplaced like 'bad' in the high ratings word cloud and 'impressed' in the low ratings. This brings up an important point in regards to the structure of english, in which positive words are often used to enhance negative sentiments and vice versa. For example, if a reviewer writes 'I wanted to see this so bad! It was awesome!' - bad is positive in this case. Or if many reviewers wrote 'I was not impressed', then impressed would show up frequently in low ratings, even though the word itself has a positive connotation or sentiment generally.

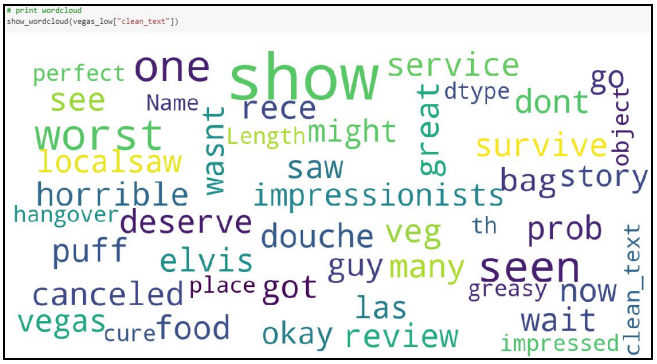


Figure 2: Performing Arts Yelp Review Word Cloud - Low Ratings (1, 2, or 3 stars)

Figure 3: Performing Arts Yelp Review Word Cloud - High Ratings (4 or 5 stars)

3.2. Methods

Our analysis consists of the following methods:

- **4.1. JiaHao.** Conduct topic modeling and feature selection to extract important topics from reviews using Latent Dirichlet Allocation and Latent Semantic Analysis.
- **4.2. Alex.** Perform Sentiment Analysis to explore how text in reviews is related to ratings.
- **4.3. Brooks.** Design a Recommender System through Collaborative filtering to suggest future businesses based on past user ratings.

4. Methods & Results

4.1. Topic Modeling - LDA and LSA Analysis JiaHao

Background

Topic modeling is a process that automatically identify topic present in a text object to derive hidden patterns exhibited by a text corpus, which assists better decision making. Topic models are useful for the purpose of document clustering, organizing large blocks of textual data, information retrieval from unstructured text, and feature selection. For our analysis, we used LDA and LSA for topic modeling.

LDA is a matrix factorization technique that assumes documents are produced from a mixture of topics, in which the topics generate words based on the probability distribution. The review text are combined to form a corpus, in which we clean and preprocess the corpus to remove the punctuations, stopwords, normalized, lemmatized, and converted into a matrix representation. LDA model looks for repeating terms patterns in the entire doc-term matrix. We used the gensim python library to handle the convert the reviews since it is scalable, robust, and efficient. We also create an object for the LDA model and train it on the matrix, where we extract the top 50 topics. The results are lines of topics with individual topic terms and weights, where we can infer insights on the key topics covered for the reviews on performing arts in Vegas. Lastly, we also create an interactive topic visualization chart that displaces topics along with the most relevant words by using the pyLDAvis library.

All languages have their own intricacies and nuances which are quite difficult for a machine to capture. This includes different words that have the same meaning, or the same words with different meanings. We as humans can easily distinguish the context behind words, but a machine would not be able to capture the concept as it cannot understand the context in which the words have been used. Since simply mapping words to documents would not help, LSA attempts to leverage the context around the words in the reviews to capture the hidden concepts or topics. In our analysis for LSA, We performed the same cleaning and preprocessing steps for the reviews and created the doc-term matrix using TfidfVectorizer from the sklearn library. Then we represent every term and document as a vector by using truncatedSVD to decompose our

matrix and print out the top 10 important words for each topic. Similar to LDA, this allows us to gain insights on important words from each topic that pertains to the performing arts businesses in Vegas.

Since the rating distribution of our data is skewed to high ratings (rating_mean = 4.25), we also performed LSA on reviews with low ratings vs. high ratings to try and identify the topics that are important. We can then infer what are the topics that users generally have high reviews on, and seek on what topics that businesses have to work on to improve their ratings that would bring benefits to their businesses.

Results

Topic Modeling - LDA and LSA Analysis

For the LDA analysis, the top 50 topics are extracted along with the weights on the topic terms. The top 5 topics and their weights are as follows:

(9, '0.064*"game" + 0.050*"machine" + 0.047*"place" + 0.031*"pinball" + 0.031*"play" + 0.017*"museum" + 0.013*"old" + 0.012*"slot" + 0.012*"hall" + 0.011*"like"')
(13, '0.049*"song" + 0.030*"music" + 0.024*"band" + 0.023*"show" + 0.019*"elvis" + 0.018*"great" + 0.018*"singing" + 0.016*"like" + 0.016*"singer" + 0.015*"fan"')
(35, '0.029*"blah" + 0.018*"dynamic" + 0.015*"jersey" + 0.014*"meh" + 0.014*"answered" + 0.013*"exhibit" + 0.013*"utterly" + 0.012*"challenging" + 0.012*"nevertheless" + 0.011*"ignore"')
(2, '0.096*"dance" + 0.087*"class" + 0.018*"instructor" + 0.017*"student" + 0.013*"environment" + 0.011*"move" + 0.010*"new" + 0.010*"absinthe" + 0.010*"week" + 0.010*"timing"')
(40, '0.039*"dinner" + 0.036*"entertainment" + 0.024*"food" + 0.019*"rabbit" + 0.018*"cocktail" + 0.018*"dessert" + 0.017*"menu" + 0.016*"live" + 0.016*"rose" + 0.016*"plate"')

Through the pyLDAvis graph, we can also see the top 30 important topics terms across all topics. From the LDA analysis, we can conclude that Yelp users are focused on giving good reviews on the shows in Las Vegas, stating that the performing art shows are great, fun, funny etc. Users also gave good reviews on magic shows, food, theaters, and music venues. Ticket prices are also one of the important topics mentioned in the reviews. Cirque du Soleil is one of the business venues that garners the most, mainly positive reviews. We can also gain different insights from different topics, where in topic 12 we can see that the negative reviews are generally focused on Yelp users wasting money on boring and disappointing shows.

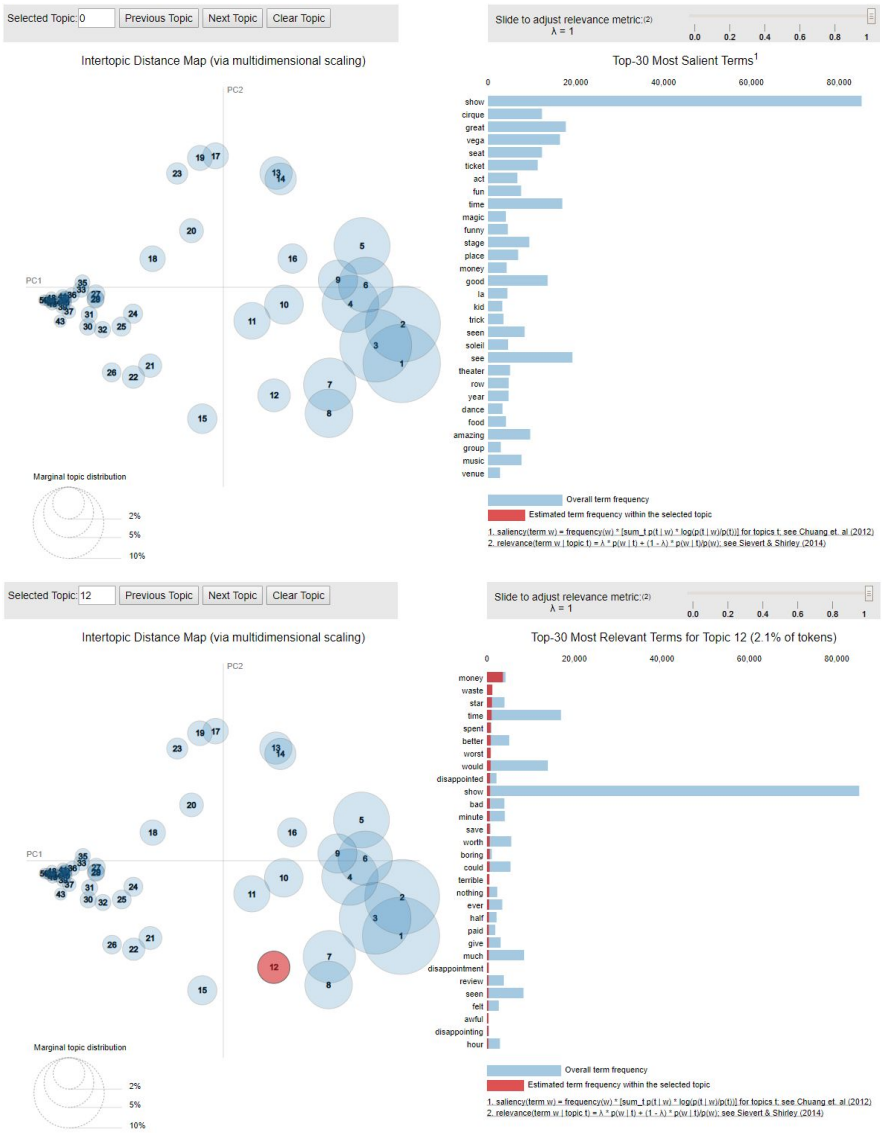


Figure 4: Interactive LDA graph showing 30 top important terms in each topic.

We can infer identical results from the LSA analysis when we do a comparison between low rating topics and high rating topics. For the topics in low rating analysis, we can identify some topics that discuss important terms from negative reviews:

Topic 4: money waste ticket cirque worst time line save soleil horrible
Topic 5: money waste time like boring music dance song guy dancer
Topic 11: vega seen like place theater game movie worst seat machine
Topic 40: david disappointed copperfield review illusion boring acrobatics dancing vega water
Topic 43: place minute horrible recommend stage watch entertaining kid waste worst

We can see from the topics in low reviews that there are some Yelp users who believes the shows to be boring and disappointing, such as David Copperfield's magic show and other musical/theater performances, and felt that money is wasted on the tickets. It is important to note

that some users are more critical in their rating, therefore even though the rating given is lower, the review may still be geared towards positive comments.

On the other hand, the LSA topics extracted from the high rating reviews show similar terms when compared to the LDA analysis. With the addition of good reviews on pinball machine games, the topics pointed out terms that garner high reviews from users, such as the Bellagio fountain, Cirque du Soleil, the Beatles music, as well as the shows effect on family time and kids:

Topic 0: time great vega cirque seat love like really music beatles
Topic 1: machine pinball game place play arcade quarter hall fame cent
Topic 2: fountain bellagio vega water free night minute time watch beautiful
Topic 3: beatles cirque love music soleil fountain song machine acrobatics water
Topic 4: magic vega time family great kid funny trick love seen

We evaluate the topics extracted from both LDA and LSA, and compare them with the actual Yelp reviews. We can see that the topics and terms gained from LDA and LSA are accurate when corresponded with the actual reviews. Therefore, we can conclude that LDA and LSA successfully provide insights on the important topics mentioned from the reviews without going through the trouble reading them individually. We can also see that people are generally happy about the performing arts businesses in Las Vegas as a whole.

Example Review with Low Rating:

Cirque du Soleil - Criss Angel Believe: “what waste of money, ours were 60.00 each and we felt ripped off !! Just a bunch of smoke and mirrors, sad to say..... I actually fell asleep !! Save your money and watch him on t.v. Criss Angel goes down 3 notches in my book. The best part was the little guy who says " foo foo foo !! " I would not have felt so bad with free tickets, and feel for those who payed more”

Example Review with High Rating:

Pinball Hall Of Fame: “Pinball machines are slowly working their way out of society, so I'm glad a place like this exists. A wonderful and loving tribute to a great past time. Lots of rare, unique, and long forgotten tables that you can play for the 1st or 100th time. There's plenty of famous tables (LOTR, Addams Family, Ripley's) that have had color conversions too. All tables are well cared for and while there's always a handful of tables that are broken they seem to actually fix them. Ample parking and low table prices. $5 or $10 bucks will get you plenty of play time. It's not walkable from the strip, but a cab or Lyft wouldn't be too expensive as you're just to the east. Highly recommended if you have even a passing interest in Pinball.”

4.2. Sentiment Analysis & Model Building
Alex

Background

Sentiment analysis is an NLP technique for analyzing text data and classifying its contents as positive, negative, or neutral. Generally, a sentiment has the attributes of Polarity, Subjectivity. For this analysis, we focus on polarity, as this is the magnitude of positive or negative sentiment on an opinion identified by sentiment analysis algorithms. Sentiment analysis is particularly valuable for applications involving large amounts of text data in which opinions are shared. Some examples of this include, analyzing survey responses, product reviews, social media comments, and emails. With this technique, a reduced metric for the contents of large body of text can be created. Sentiment is usually reported and quantified on a scale from -1 to 1 or 0 to 1, with 1 being the most positive and -1 or 0 being most negative.

By applying sentiment analysis to the Yelp Review data we demonstrate the utility of this technique in providing a quick understanding of the contents of reviews.

Results

Part 1: Yelp Reviews & Sentiment Analysis with TextBlob

1. **Get the data**
2. **Select just the text and review_stars data**
3. **Add a column for highlow, where high is 4 or 5 stars and low is 1,2, or 3 stars**
4. **Clean the text column**
5. **Analyze sentiment using the textblob package**

Sentiment is given as a score between -1 and +1. The more negative, the greater the probability that the sentiment was poor - The more positive, the more likely it was a good sentiment in the text. Use of this package was to establish a baseline in analyzing sentiment.

```
#vegas_high (4 and 5 star ratings)
vegas_high = vegas[vegas['highlow'] == 1]
vegas_high = vegas_high.sort_values(by=['textblob_sentiment'], ascending=False)
vegas_high.head()
```

	text	review_stars	highlow	clean_text	textblob_sentiment
27994	I'm speechless!! Can't describe how wonderful ...	5.0	1	im speechless cant describe how wonderful it i...	1.0
19802	Absolutely awesome show!	5.0	1	absolutely awesome show	1.0

```
#vegas_low (1,2,3 star ratings)
vegas_low = vegas[vegas['highlow'] == 0]
vegas_low = vegas_low.sort_values(by=['textblob_sentiment'])
vegas_low.head()
```

	text	review_stars	highlow	clean_text	textblob_sentiment
30149	Show got canceled and the service was horrible...	1.0	0	show got canceled and the service was horrible...	-1.0
20341	It was the worst show we have ever seen; there...	1.0	0	it was the worst show we have ever seen there ...	-1.0

As a quick sanity check, note the contents of the clean_text column and the associated textblob_sentiment score. With a clear positive score of 1.0 we see a review stating that a user was speechless and that the show was absolutely awesome. With a clear negative score of -1.0 we see a review stating that a user's show was cancelled and another that it was the worst show ever seen. This tells us that sentiment can indeed be readily gleaned from the text content of Yelp Reviews.

6. **Get high and low dataframes, and visualize frequent words in a word cloud**
7. **Visualize sentiment scores vs. ratings and sentiment scores vs. highlow**

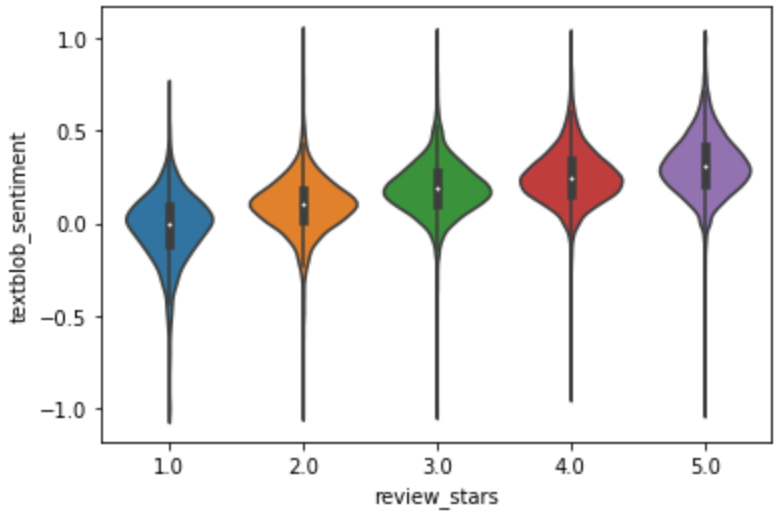


Figure 5: TextBlob_Sentiment vs ReviewStars Rating with Correlation = 0.55

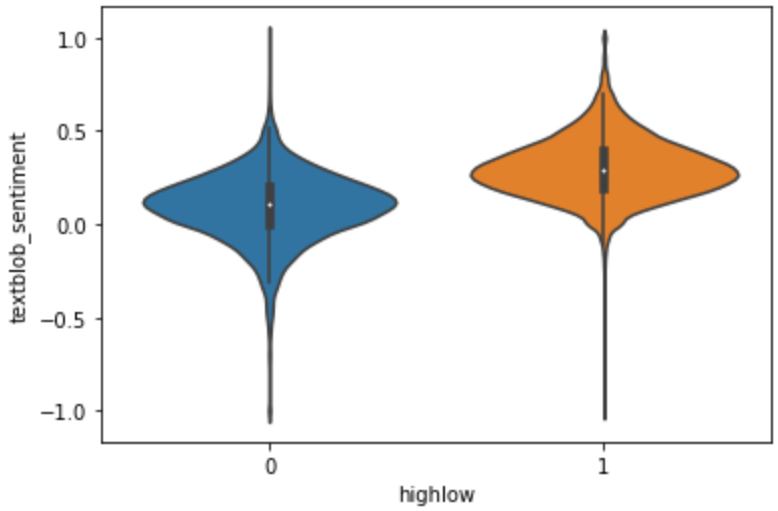


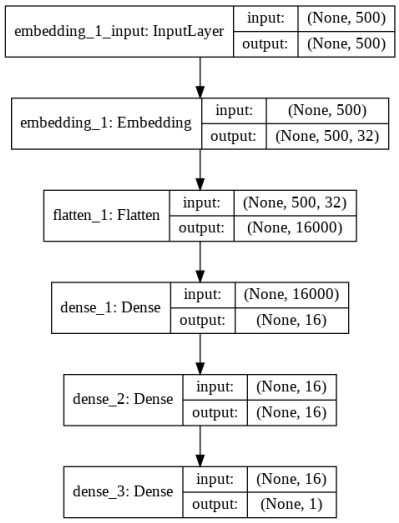
Figure 6: TextBlob_Sentiment vs HighLow Rating with Correlation = 0.48

From the above chart, we can tell that the textblob sentiment is positively correlated with the review stars that were given. High sentiment in a review tends to occur alongside high ratings and the same for low sentiment and low ratings.

Part 2: Train Model on classic IMDB movie review dataset, and use the model to analyze sentiment on yelp reviews.

We really wanted to know the extent to which an outside model could apply to Yelp reviews, assuming language is similar among reviews in different fields. We were impressed and excited to find that it does work to train a model on one dataset, and apply it to a novel dataset. We used the IMDB movie review dataset and applied the sentiment model learned there to our Yelp review dataset.

1. **Get IMDB dataset which contains movie reviews data and preprocess it**
2. **Using a standard sequential model, relu activations, adam optimizer, and a binary cross entropy loss function**
3. **With a batch size of 128 and 5 epochs, train the model**




```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 500, 32)           320000
_________________________________________________________________
flatten_1 (Flatten)          (None, 16000)             0
_________________________________________________________________
dense_1 (Dense)              (None, 16)                256016
_________________________________________________________________
dense_2 (Dense)              (None, 16)                272
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 17
=================================================================
Total params: 576,305
Trainable params: 576,305
Non-trainable params: 0
```

Figure 7: Multilayer Dense Sequential Model for learning sentiment in IMDB movie reviews using the Keras package.

4. **Display the results of training**

```
25000/25000 [==============================] - 7s 292us/step - loss: 0.5289 - acc: 0.6988 - val_loss: 0.3205 - val_acc: 0.8614
Epoch 2/5
25000/25000 [==============================] - 6s 248us/step - loss: 0.1949 - acc: 0.9272 - val_loss: 0.2834 - val_acc: 0.8830
Epoch 3/5
25000/25000 [==============================] - 6s 251us/step - loss: 0.0703 - acc: 0.9803 - val_loss: 0.3516 - val_acc: 0.8731
Epoch 4/5
25000/25000 [==============================] - 6s 252us/step - loss: 0.0174 - acc: 0.9972 - val_loss: 0.4365 - val_acc: 0.8718
Epoch 5/5
25000/25000 [==============================] - 6s 256us/step - loss: 0.0050 - acc: 0.9995 - val_loss: 0.4764 - val_acc: 0.8712
```

Figure 8: Shows results of training in each epoch. Final model accuracy of 87.12%

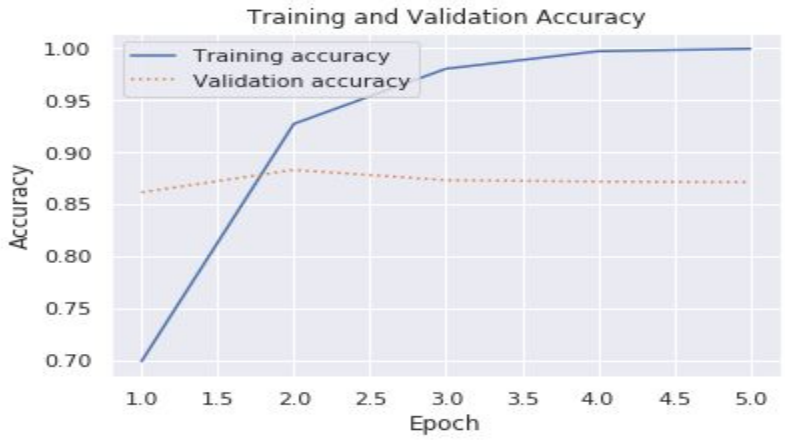


Figure 9: This figure visualizes accuracy on the training and validation throughout the training process. Some overfitting is likely given our large train/validation accuracy split, but it is sufficient for our exploratory purposes.

5. **Build an analyzer function to apply this outside model to our Yelp data**
6. **Visualize the new model's sentiment scores vs ratings**

	text	review_stars	highlow	clean_text	textblob_sentiment	imdb_model_sentiment
0	Like walking back in time, every Saturday morn...	4.0	1	like walking back in time every saturday morni...	0.061458	0.999334
1	Glad we caught it before leaving the Vegas sta...	5.0	1	glad we caught it before leaving the vegas sta...	0.292857	0.999253
2	This is a really quaint place for the whole fa...	4.0	1	this is a really quaint place for the whole fa...	0.177941	0.999648
3	Yes! As good as it gets in way of pinball mach...	5.0	1	yes as good as it gets in way of pinball machi...	0.448889	0.999035
4	My family and I have been to this show twice. ...	5.0	1	my family and i have been to this show twice b...	0.290747	0.999693

Figure 10: First 5 reviews and associated sentiment scores from our 2 models.

When we apply this new model to our Yelp dataset, it becomes clear right off the bat that it is more effectively classifying sentiment than TextBlob was. Reading the first 5 reviews above, we see that textblob_sentiment was mischaracterizing the sentiment of positive reviews with low scores, which does not appear to be the case with the imdb_model_sentiment scores.

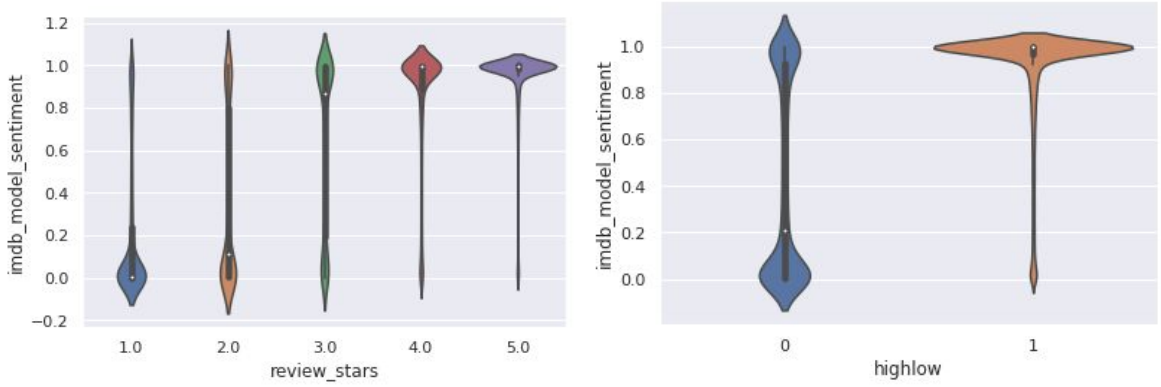


Figure 11: The relationship between sentiment and review ratings is visualized above. Correlations between imdb_mode_sentiment and review_stars is 0.65 and between imdb_model_sentiment and highlow is 0.6. These are higher than for textblob and groups appear much more distinct.

7. Test the model on some Yelp Reviews

```
positive_review_example = vegas_high.iloc[0]['clean_text']
positive_review_example
```

```
'im speechless cant describe how wonderful it is and how impressive ive got i will bring my family to watch the show
```

```
analyzer(positive_review_example)
```

```
0.9858878
```

A sentiment score of 0.986 is high, and this is correct as the contents of that review were positive!

```
negative_review_example = vegas_low.iloc[0]['clean_text']
negative_review_example
```

```
'show got canceled and the service was horrible about getting refund hoping i can get a refund tomorrow'
```

```
analyzer(negative_review_example)
```

```
0.096437775
```

A sentiment score of 0.096 is low, and this is correct as the contents of that review were negative! From this I can conclude that it is in fact possible to leverage existing sentiment models, trained on other datasets for the task of analyzing sentiment in reviews. At least in an exploratory capacity, it seems that this technique is effective.

8. Extra - Predictive ability check - Logistic Regression on High-Low Ratings

In this last step I trained a logistic regression model on the Yelp dataset using the single sentiment analysis score provided by the imdb model to predict whether a rating would be high or low. For train/test split I used 80% train to %20 test. Results are shown below:

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=None, solver='warn', tol=0.0001, verbose=0,
                   warm_start=False)
```

```
# simple accuracy check
score = logreg.score(xr_test, yr_test)
print(score)
```

```
0.8385126162018592
```

This suggests that we can predict with about 83.8% accuracy whether the star rating of a review is high (4 | 5) or low (1 | 2 | 3) based on the text of the review. This suggests that the sentiment analysis using sequential model in keras on imdb data can effectively analyze sentiment in text. Not only that but we can extrapolate this model to our own dataset and accurately use it to predict whether a rating will be high or low. Sentiment can therefore serve as a reduced representation of the text data in model building for Yelp Reviews. Importantly this can then be related back to the rating a customer left.

4.3. Recommender Systems
Brooks

Background

There are several different Recommender Systems techniques which can be effectively utilized in the building of user recommendations. These techniques include rule-based, content/feature based, and collaborative filtering. Each of these techniques can be used in building recommendations for specific data sets or in tandem with one another.

Rule-based recommendations are generally used as a starting point when data about a user's preferences are not yet known. This can often occur when a user signs up for a service, like Yelp, but has yet to submit enough reviews to build an in-depth understanding of their preferences. In rule-based recommendations data about the user's demographic, geographic, and psychographic information is used to provide suggestions for the user. For instance, if user A has the same age, gender, geographic location, and marital status as user B, then the algorithm will assume user A will generally like the same things as user B. In practice, there will be thousands of users like user A which helps to make rule-based recommendations more robust and a good starting point for building recommendations for new users.

Content, or feature, based recommendations is a more in-depth approach to building recommendations for users. The technique examines the different features which make up products, services, or businesses a user enjoys and then searches for comparable products/services/businesses with similar features. For instance, if a user generally rates Japanese restaurants or green colored clothing highly, the feature-based approach to recommender systems will suggest other Japanese restaurants and green colored clothing to the user. This technique requires an in-depth understanding of the features which make up a product/service/business as well as frequent user interactions with each of the features. Using clothing as an example, there are several features which could be important for defining the clothing. Features of interest might be the color, the size, the brand, the style, and even the seasonal component of when the item is most often purchased. The user will need to have interacted with each of these features as well, having often purchased items of the same color, brand, size, and style. The content-based approach can be highly customized and personalized to each user, but requires a granular level of information about the both the user and the area of focus for the recommendation.

Collaborative filtering is an approach to recommendations which examines the relationship between users in terms of their propensity for rating items similarly. In this technique groups of users who rate several products/services/businesses highly are assumed to continue to do so for other products/services/businesses that they might not have interacted with yet. A recommendation can be suggested based on this assumption. For example, if user A is clustered into group Z based on the fact that user A generally rates the same things highly as other members of group Z, then it can be assumed that user A will enjoy the other things that members of group Z have rated highly. If user A has not yet interacted with product X and members of the group Z, which user A belongs, have rated product X highly, then user A will be recommended product X. Other approaches to collaborative filtering can be utilized, such as finding a person

who rates things most similarly to user A and providing user A with recommendations based on the user with ratings most similar to them.

Each of these techniques can be boiled down to the following. Rule-based recommendation is founded on who the user is, content-based recommendations are founded on the features that make up the items they enjoy, and collaborative filtering is based on the idea a person's habits are similar to the habits of someone else.

Methods

For this analysis the Collaborative Filtering approach to Recommender Systems will be applied to a Yelp reviews dataset in order to yield powerful recommendations to users for businesses to visit in the Las Vegas area. Only users and businesses with considerably large number of reviews will be utilized, so a starter recommendation algorithm founded on the rule-based approach for recommendation will not be necessary. Furthermore, the content-based approach will not be utilized as features concerning each business will not be incorporated in the analysis. Dimensionality reduction through Singular Value Decomposition (SVD) will be applied to handle the sparsity of the data and several approaches to the Collaborative Filtering technique will be examined and measured in terms of accuracy. In addition, grid-based parameter selection will be applied to the best performing Collaborative Filtering modeling technique in order to produce the best possible algorithm for providing recommendations. Finally, example recommendations will be produced and analyzed.

Data & Preprocessing for Recommender System

Only users and business with numerous reviews were utilized in this portion of the analysis. Thus, any user with less than 10 reviews and any business with less than 10 reviews were removed from the data set. Also, all businesses across each category were included. The final data which was used in this portion of the analysis contained 7,825 reviews, by 251 users, for 278 businesses. Thus the recommender system specifically focused on providing reviews for 251 individuals based on their previous business ratings. Each user_id and business_id contained in the data set were reset to reflect a 0 to n value.

(7825, 3)

	user_id	business_id	stars
0	144	0	5.0
1	144	1	5.0
2	144	2	4.0
3	144	3	5.0
4	144	4	4.0
5	144	5	5.0
6	144	6	1.0
7	144	7	1.0
8	144	8	5.0
9	144	9	5.0

Figure 12: Yelp users and associated business id and star ratings given.

SVD

The first technique applied on the data was Singular Value Decomposition (SVD). This was done to both reduce the dimensionality of the data and to produce predicted recommendations. The first step in performing SVD is to convert the data set into a matrix form with the rows representing the user_id, the columns representing the business_id, and the values within the matrix being the crosstabulation of star ratings given by the user to the business. Even with the processing of the data set to include only users and businesses with at least 10 reviews, the data set is still rather sparse. In all there are 251 users and 278 businesses, making for a matrix with 69,250 possible values (251 x 278). However, there are only 7,825 total values present in the matrix. Keep in mind this is a crosstabulation of user ratings of businesses. This means that most users have rated a small portion of the businesses and most businesses have been rated by a small portion of users. Without reducing the data matrix into a small dimensional space, it would be difficult to provide recommendations for users.

business_id	0	1	2	3	4	5	6	7	8	9	...	268	269	270	271	272	273	274	275	276	277
user_id																					
0	3.5	4.0	0	0.0	0	4.333333	0	0	4.0	3.5	...	0.0	4.0	4	0.0	0	0	0.000000	4	0.0	3.0
1	4.0	4.0	0	0.0	5	5.000000	0	2	0.0	4.0	...	0.0	0.0	0	5.0	0	2	0.000000	5	4.0	0.0
2	4.0	5.0	2	0.0	0	4.000000	4	5	4.5	5.0	...	3.0	0.0	0	5.0	0	5	4.000000	0	0.0	4.0
3	0.0	0.0	0	0.0	3	3.000000	0	0	0.0	5.0	...	4.0	0.0	0	0.0	0	0	4.666667	0	0.0	0.0
4	5.0	5.0	0	0.0	5	0.000000	4	0	0.0	4.0	...	0.0	0.0	4	0.0	0	0	0.000000	4	4.0	4.0

Figure 13: Matrix of user ratings of each business. Note sparsity given that users have rated only a small number of the total possible businesses.

Singular Value Decomposition (SVD) was performed to reduce the dimensionality of the data matrix in order to control for sparseness within the data set. As previously mentioned, the overall number of dimensions in the data set is 278 (based on the number of businesses), thus any reduction technique would need to reduce the dimensions down to a small feature space than

278. The SVD approach also works best on data that is normalized across users. Each user has a different rating tendency, some users will never rate a business as higher than 4 or lower than 2. This means the best possible rating they would give to a business is 4. Other users might frequently rate businesses as high as 5 and 4. Meaning the lowest rating they ever give is 4. When this occurs, users are using a subjective rating scale and one user's rating of 1 might be the same as another user's rating of 4. Prior to utilizing the SVD technique the user ratings were placed on a scale of (-1,1) based on their own ratings tendency.

Upon the application of SVD to the user review data set a total of 50 underlying concepts were identified. Inherently SVD tries to combine the data into underlying latent concepts based on the idea that similar things (in this case businesses) perform similarly to one another resulting in the combining of the similar businesses into a single feature. The application of SVD suggested there were a total of 50 different business concepts which users based their ratings. The 50 concepts can be identified through the examination of the sigma output matrix and the values of the diagonals from the SVD technique.

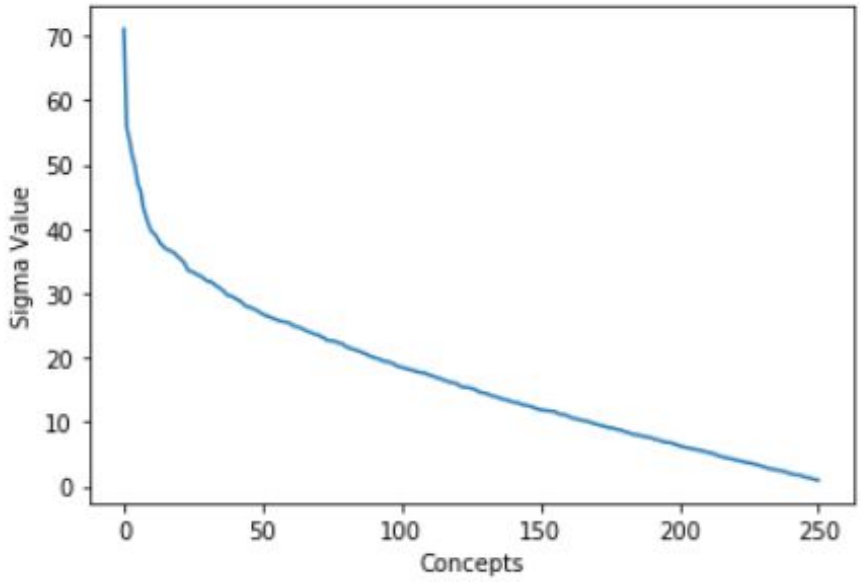


Figure 14: Sigma Value vs Concepts using SVD

```
[[71.09083129  0.          0.          ...  0.          0.
   0.        ]
 [ 0.         55.76264303  0.          ...  0.          0.
   0.        ]
 [ 0.          0.         53.86397254 ...  0.          0.
   0.        ]
 ...
 [ 0.          0.          0.         ...  1.1089647   0.
   0.        ]
 [ 0.          0.          0.         ...  0.          1.01164499
   0.        ]
 [ 0.          0.          0.         ...  0.          0.
   0.92135633]]
```

Figure 15: Reduced feature space with SVD.

In addition to reducing the feature space, SVD can be used to predict and provide a recommendation to a user based on their affinity for the concepts SVD has identified. Basically, if a user rates a Japanese restaurant highly and the restaurant is captured with in one of the concepts identified by SVD, then the user will likely rate other businesses within that concept high as well.

To demonstrate how SVD can be utilized to provide a recommendation based on the identification of latent concepts, the SVD algorithm for recommendation was utilized on user id 55 to produce a recommendation of 5 new businesses.

The current business the user 55 has visited and provide top star recommendations is as follows:

	user_id	business_id	stars	name
3	55	37	5.0	John Mull's Meats & Road Kill Grill
5	55	200	5.0	Black Bear Diner
10	55	156	5.0	Eat.
11	55	61	5.0	Leticia's Mexican Cocina
4	55	262	4.5	Jamms Restaurant

Figure 16: User 55 activity and ratings of restaurants.

This suggests the user likes American Western cuisine and primarily utilized Yelp to review food restaurants. The data set includes all businesses in the Las Vegas area, including casinos, shows, tourist attractions, and shopping malls. Thus, an accurate recommendation should include only restaurants of the American Western cuisine variety.

The recommendation of 5 businesses user 55 has not visited based on the SVD approach is as follows:

	business_id	name
5	7	Hash House A Go Go
56	59	Big Dog's Draft House
144	151	Viva Las Arepas
112	118	Strip N Dip Chicken Strips
29	31	BabyStacks Cafe

Figure 17: Recommendation of 5 businesses for User 55 to try out based on past ratings.

The output recommendations seem to be well aligned with the highly rated businesses and business concepts the SVD technique identified for user 55. All recommendations are for restaurants and appear to be similar to the restaurant the user previously rated highly.

Collaborative Filtering Algorithms

Once the power of the SVD approach is demonstrated, in terms of dimension reduction, concept identification, and recommended prediction, it is important to analyze several additional approaches to Collaborative Filtering in hopes of identifying the best recommender system. In this portion of the analysis several techniques are performed on the user reviews data set and analyzed in terms of cross-validation accuracy.

Types:

- **Normal Predictor** – random rating based on the distribution of the data set
- **Base Line** – estimates based on the baseline the user rates
- **KNN** – utilizes the Nearest Neighbor approach by recommending based on the most similar user to the user of interest
- **KNN Means** – the same as KNN except the users ratings habits are analyzed and ratings are normalized based on MinMax (user only rates between 4-5 or 1-3)
- **KNN ZScore** – the same as KNN except the users ratings habits are analyzed and ratings are normalized based on ZScore (user only rates between 4-5 or 1-3)
- **SVD** – Identification of concepts recommendations are made based on the other items within the concepts
- **SVDpp** – the same as SVD except the user ratings habits are taken into account (user only rates between 4-5 or 1-3)
- **CoClustering** – utilizes clustering techniques to group similar items and users together
- **SlopeNorm** – utilizes portioning via lines to separate groups of items and users

Algorithm	test_rmse	fit_time	test_time
SVDpp	0.901591	1.957468	0.078719
SVD	0.909810	0.235718	0.013501
KNNBaseline	0.919262	0.009433	0.113966
KNNWithMeans	0.921412	0.013599	0.071066
KNNWithZScore	0.934118	0.010545	0.092044
SlopeOne	0.944042	0.008523	0.051268
KNNBasic	0.958206	0.011136	0.062663
CoClustering	0.960721	0.108050	0.013128
NMF	1.009864	0.244940	0.010210
NormalPredictor	1.306646	0.000705	0.012863

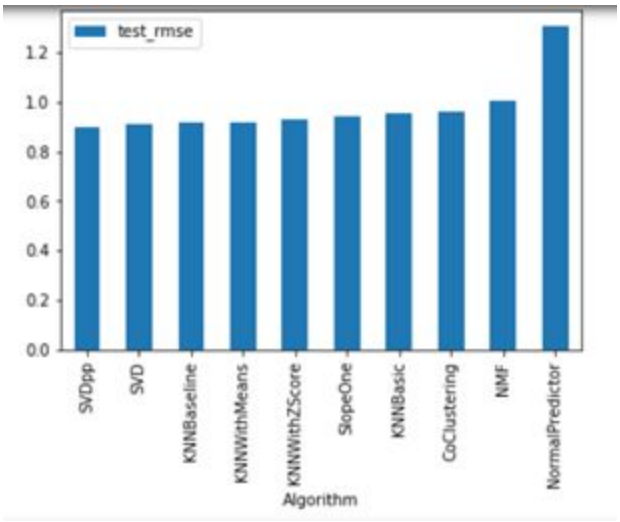


Figure 18: Model Selection identifies the SVDpp algorithm for best recommendation power.

The results suggest that the SVDpp algorithm is the most accurate in terms of prediction, given the test_rmse, but is also the most computationally expensive in terms of time to process. Once the SVDpp algorithm was identified as the most accurate technique to apply, it was tested on user 55 to find the recommendation the algorithm produces for the user.

Recommendation:

```
Top item for user 55 is  Yonaka Modern Japanese with predicted rating [4.87892285]
```

As previously stated, this user tends to enjoy American Western Cuisine thus the output of the model for this user seems to be a little more inaccurate than the previously utilized SVD model.

Parameter Tuning

In addition to identifying the best algorithm, the parameters of the algorithm can also be adjusted to find the best combination. This can be accomplished through a grid based approach to parameter tuning. Since the SVDpp algorithm was identified as being the best recommendation predictor, in terms of its base parameters, a grid-based approach to parameter tuning was utilized on the SVDpp parameters.

The primary parameters of interest were the Learning Rate, Regulation Term, and Number of Factors. As previously stated, the original SVD analysis yielded results suggestion 50 total concepts, and the Factor parameter takes as an input the number of concepts expected in the data. The default for this parameter is 20 and additional values will be tested, including 10, 20, 50, 10.

The output of the grid base parameter tuning approach yielded the following suggest parameters: Learning Rate = 0.01, Regulation Term = 0.5, Number of Factor = 10. The SVDpp

model was updated with the new values for each of the parameters and cross-validation was performed on the reviews data set across 5 folds.

```
Evaluating RMSE, MAE of algorithm SVDpp on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9137  0.9011  0.9159  0.8658  0.8819  0.8957  0.0192
MAE (testset)     0.7065  0.7049  0.7024  0.6608  0.6850  0.6919  0.0174
Fit time          2.14    2.13    2.09    2.03    2.05    2.09    0.04
Test time         0.07    0.05    0.06    0.05    0.05    0.06    0.01
```

Figure 19: Best Recommender Model Evaluation.

The output produced an average RMSE of 0.8957 across each of the 5 folds. This is an improvement from the original base SVDpp model with no parameter which yielded a RMSE of 0.90159, suggesting the parameter tuned SVDpp model is the best model for producing recommendations. The final step was to apply the parameter adjusted model to the user 55 to

Recommendation:

```
Top item for user 55 is Yardbird Southern Table & Bar with predicted rating [4.61528523]
```

The new recommendation is much more aligned with the rating habits of user 55 as it appears to be more along the lines of American Western cuisine, which the user had previously rated highly.

Conclusion

The results of the Recommender System through Collaborative filtering help show how different recommender algorithm techniques can be utilized on sparse data sets to produce accurate recommendations. Obviously, there is still concerns with computational cost for large scale recommender systems, but in terms of accuracy the SVDpp algorithm with parameters Learning Rate = 0.01, Regulation Term = 0.5, and Number of Factor = 10 is suggested for the Las Vegas users and the business portion of the Yelp data set.

5. Discussion & Future Work

5.1. Discussion

For the Topic Modeling using LDA and LSA portion, we conducted the analysis across all reviews for the performing arts businesses in Las Vegas. While we prove that we can garner both positive and negative insights from the reviews for all businesses as a whole, the topics are still in a very broad spectrum. Ideally, the analysis should be done for each individual business to find out the positive and negative terms mentioned for the specific business so that they know their own strengths and weaknesses to improve. Aside from that, we can further improve topic modeling results by using frequency filters, part of speech tag filters, and batch wise LDA. In the future, we would also like to perform feature selection in the future to identify the importance of the topics modeled against the review rating regarding the businesses to strengthen the conclusion of the LDA/LSA analysis. It is important to point out that different users have different rating scales for their reviews. Additionally, people gave higher reviews based on environmental factors, since Las Vegas is generally a place with fun and great entertainment. Therefore, the reviews should be normalized against the user's information, such as average review ratings and review count to give a more accurate analysis for the topic models.

Sentiment analysis provided us insights into the relationship between the text left in Yelp Reviews and ratings users gave. We tested two models of sentiment analysis and both produced scores that were moderately correlated with review stars and high/low ratings. The simple sentiment model using the textblob package was less effective than the model using a multilayer sequential model in keras built off of the IMDB dataset. We also found that interestingly enough, training a sentiment model on a sufficiently large outside dataset of reviews like IMDB is enough to analyze sentiment on a new dataset like our Yelp Reviews one. A simple logistic regression model to predict between high and low ratings on sentiment alone was 83.8% accurate, further supporting the effectiveness of this method. It suggests that sentiment score is an efficient means of reducing a body of review text into a useable metric for modeling.

The Recommender System analysis allowed us to examine various techniques for building recommendation algorithms as well as examining how robust statistical techniques can be effectively utilized on sparse data. Of note was how similarly each of the algorithms performed in producing accurate recommendations. The only algorithm which performed relatively poorly was the NormalPredictor, which relies solely on the distribution of the data. The first technique applied, which was the SVD algorithm, performed really well on its own. This suggests that the reduction in feature space through SVD can be highly powerful on its own. Although computational cost was examined in the analysis, it was never considered a factor in terms of producing the best algorithm. If the analysis was built with the idea of creating a productionalized algorithm for ongoing recommendations in a larger feature space, this aspect would need to be considered.

5.2. Future work

The Recommender System which was built for this analysis made use of the Collaborative Filtering technique to generate suggestions for the users. As previously stated, only users and businesses with several reviews were utilized in the building of the algorithm. A future analysis should examine how rule-based recommender systems could be built for users who are new to the Yelp service. This type of recommender system could take into account geographic, demographic, and psychographic details about the user in order to generate an initial recommendation. On the other side, a more specific and personalized recommendation system could be built through the application of the content-based approach. The Yelp dataset contains metadata about each of the businesses, such as category of business, number of reviews, and location. Also, the wording of the reviews themselves could be used to further categorize the businesses. Similarly to what was done in the sentiment analysis portion of this article. These additional features would allow for a content-based recommender system to be built with the hopes of providing any even more accurate and personalized recommendation.

APPENDICES

LDA Analysis: Topic models and weights

```
[(23, '0.047*"ambiance" + 0.044*"popcorn" + 0.042*"card" + 0.028*"credit" + 0.019*"eating" + 0.019*"movie" + 0.017*"cash" + 0.013*"sharing" + 0.013*"news" + 0.013*"bmg"'),
(5, '0.116*"show" + 0.033*"see" + 0.024*"vega" + 0.021*"time" + 0.021*"great" + 0.021*"amazing" + 0.017*"it" + 0.014*"would" + 0.014*"loved" + 0.013*"recommend"'), (41, '0.0
35*"tap" + 0.022*"notch" + 0.021*"top" + 0.016*"hip" + 0.016*"study" + 0.015*"rrl" + 0.014*"champagne" + 0.014*"octopus" + 0.013*"lobster" + 0.011*"celebration"'), (44, '0.0
25*"volunteer" + 0.022*"terry" + 0.017*"skeptical" + 0.016*"puppet" + 0.014*"talent" + 0.014*"stage" + 0.013*"fator" + 0.013*"he" + 0.011*"hypnosis" + 0.011*"kevin"'), (9,
'0.064*"game" + 0.050*"machine" + 0.047*"place" + 0.031*"pinball" + 0.031*"play" + 0.017*"museum" + 0.013*"old" + 0.012*"slot" + 0.012*"hall" + 0.011*"like"'), (48, '0.022
*"dean" + 0.017*"coach" + 0.010*"supper" + 0.009*"donnie" + 0.008*"belly" + 0.008*"mussel" + 0.008*"pho" + 0.006*"trilogy" + 0.005*"sport" + 0.005*"preparation"'), (6, '0.02
6*"house" + 0.021*"line" + 0.017*"experience" + 0.017*"zombie" + 0.014*"haunted" + 0.013*"actor" + 0.010*"wait" + 0.009*"time" + 0.009*"ride" + 0.009*"get"'), (47, '0.028*"b
ill" + 0.017*"racist" + 0.016*"supposedly" + 0.014*"kicked" + 0.013*"refused" + 0.013*"bank" + 0.012*"guaranteed" + 0.011*"ugh" + 0.010*"dollar" + 0.010*"expert"'), (37, '0.
049*"center" + 0.035*"balcony" + 0.031*"smith" + 0.025*"self" + 0.024*"season" + 0.022*"building" + 0.019*"parking" + 0.019*"breakfast" + 0.019*"hall" + 0.015*"art"'), (4,
'0.022*"stage" + 0.021*"amazing" + 0.020*"show" + 0.019*"performer" + 0.017*"performance" + 0.016*"story" + 0.015*"water" + 0.012*"music" + 0.012*"scene" + 0.011*"beautifu
l"'), (1, '0.021*"listen" + 0.018*"thursday" + 0.017*"partner" + 0.017*"response" + 0.015*"refund" + 0.014*"dramatic" + 0.013*"floating" + 0.013*"vegascom" + 0.012*"overwhel
ming" + 0.012*"ending"'), (18, '0.065*"food" + 0.041*"place" + 0.028*"good" + 0.022*"dish" + 0.019*"burger" + 0.016*"ordered" + 0.016*"eat" + 0.015*"restaurant" + 0.013*"ser
vice" + 0.013*"experience"'), (3, '0.126*"group" + 0.104*"blue" + 0.095*"man" + 0.024*"paper" + 0.023*"men" + 0.019*"paint" + 0.014*"paintball" + 0.012*"chef" + 0.012*"toile
t" + 0.012*"field"'), (22, '0.032*"fresh" + 0.022*"sooo" + 0.019*"bday" + 0.017*"strongly" + 0.016*"intended" + 0.016*"electric" + 0.015*"creating" + 0.014*"earned" + 0.013
*"skeptic" + 0.013*"yr"'), (49, '0.038*"phone" + 0.026*"camera" + 0.022*"security" + 0.020*"picture" + 0.019*"allowed" + 0.019*"dog" + 0.019*"ate" + 0.019*"arena" + 0.014*"f
lash" + 0.013*"horrible"'), (46, '0.118*"kid" + 0.065*"family" + 0.063*"old" + 0.055*"year" + 0.035*"child" + 0.035*"friendly" + 0.027*"great" + 0.026*"son" + 0.025*"adult"
+ 0.019*"age"'), (0, '0.050*"loud" + 0.033*"ear" + 0.031*"drum" + 0.021*"instrument" + 0.021*"thrown" + 0.019*"noise" + 0.016*"hat" + 0.013*"percussion" + 0.013*"ring" + 0.0
10*"extraordinary"'), (38, '0.135*"meet" + 0.107*"daughter" + 0.061*"greet" + 0.021*"sin" + 0.019*"ballet" + 0.017*"city" + 0.012*"understanding" + 0.011*"meeting" + 0.011
*"sexually" + 0.010*"foster"'), (8, '0.120*"cirque" + 0.108*"show" + 0.045*"soleil" + 0.038*"seen" + 0.027*"one" + 0.021*"ive" + 0.020*"love" + 0.017*"acrobatics" + 0.014*"l
ike" + 0.011*"see"'), (35, '0.029*"blah" + 0.018*"dynamic" + 0.015*"jersey" + 0.014*"meh" + 0.014*"answered" + 0.013*"exhibit" + 0.013*"utterly" + 0.012*"challenging" + 0.01
2*"nevertheless" + 0.011*"ignore"')]
```

LSA Analysis: Top 50 important topics across all reviews

```
Topic 0: great time vega cirque seat like really good amazing ticket
Topic 1: cirque soleil seen beatles acrobatics amazing stage music love story
Topic 2: machine pinball game cirque place soleil love beatles play arcade
Topic 3: great amazing vega music loved love best recommend highly time
Topic 4: magic trick audience seen vega funny criss machine pinball time
Topic 5: vega fountain best time food seen burger bellagio water free
Topic 6: seat ticket vega fountain best theater water bellagio free view
Topic 7: music song like love fountain really beatles dance dancing guy
Topic 8: amazing stage loved recommend performer really definitely absolutely highly audience
Topic 9: magic love trick beatles seat amazing penn music teller criss
Topic 10: fountain water great bellagio really beautiful magic trick free watch
Topic 11: good vega best really entertaining funny recommend seen enjoyed seat
Topic 12: time seat seen best good stage theater venue house like
Topic 13: time good love fountain amazing loved really seat recommend definitely
Topic 14: amazing worth money good really music price best ticket song
Topic 15: really enjoyed vega beatles seat like going great house cool
Topic 16: loved funny best seen kid love great water year absolutely
Topic 17: money worth seat recommend definitely awesome loved highly best burger
Topic 18: loved music year family recommend highly kid song time magic
Topic 19: vega like money performer audience funny performance theater burger stage
Topic 20: vega definitely guy cirque soleil seat girl awesome good night
Topic 21: like amazing recommend highly kid money good year class love
Topic 22: like beatles funny amazing song night definitely carrot people great
Topic 23: funny money night venue really waste drink criss cirque people
Topic 24: definitely kid family awesome worth people music line experience beatles
Topic 25: beatles people vega story burger music good money great line
Topic 26: awesome line story love stage good funny people night went
Topic 27: awesome burger dance amazing ticket seat kid michael funny free
Topic 28: music dancing dance vega theater acrobatics funny girl time beautiful
Topic 29: line entertaining seen people elvis song story theater burger definitely
Topic 30: year seen better night definitely burger seat went think criss
Topic 31: entertaining beatles song family stage awesome theater friendly girl enjoyed
Topic 32: entertaining audience like blue group guy love seen night participation
Topic 33: definitely entertaining best like line seat love magic criss michael
Topic 34: michael jackson elvis stage worth funny went penn teller night
Topic 35: year experience entertaining worth penn funny teller performance guy like
Topic 36: theater night burger went audience enjoyed money performer awesome friend
Topic 37: theater entertaining place music stage criss night angel performance went
Topic 38: performance line story funny like guy money place night better
Topic 39: place people act performer night seat talented beautiful penn girl
Topic 40: definitely year water performer funny money guy music ticket penn
Topic 41: performance penn people watch enjoyed performer teller water theater michael
Topic 42: seen experience michael jackson family performance people stage fountain beautiful
Topic 43: performance guy venue burger water people worth family drink criss
Topic 44: elvis dance enjoyed people beautiful audience performance criss year water
Topic 45: elvis star experience act family friend beautiful music little guy
Topic 46: watch song enjoyed price food beautiful worth favorite watching band
Topic 47: people funny better star family theater enjoyed talented fountain criss
Topic 48: star dance drink seat favorite food come comedy class minute
Topic 49: better water night elvis watch make beautiful reve want place
```

LSA Analysis: Top 50 important topics across low rating reviews:

Topic 0: cirque like time good really ticket seen vega seat money
Topic 1: cirque soleil seen acrobatics criss beatles love angel act story
Topic 2: trick magic criss money angel ticket waste time illusion audience
Topic 3: burger food magic criss trick angel cirque fry place money
Topic 4: money waste ticket cirque worst time line save soleil horrible
Topic 5: money waste time like boring music dance song guy dancer
Topic 6: ticket good price worth burger funny act free boring entertaining
Topic 7: seat stage trick story theater line burger view seating seen
Topic 8: good seat music beatles criss money great song angel elvis
Topic 9: trick magic food music service beatles great line love time
Topic 10: beatles love burger like ticket music elvis trick song fry
Topic 11: vega seen like place theater game movie worst seat machine
Topic 12: good really like place game pretty kid time machine line
Topic 13: cirque trick good soleil like girl drink guy people magic
Topic 14: good line burger seen vega time worst theater year song
Topic 15: game time machine place pinball burger kid good seat audience
Topic 16: really time seat vega water asleep going beatles thought thing
Topic 17: really line girl seen story guy magic dancer worst dance
Topic 18: time great like felt act soleil year kid cirque ticket
Topic 19: theater really time movie elvis vega waste great song cirque
Topic 20: star act really great review elvis performer ticket performance service
Topic 21: great funny people really act seen beatles line worst joke
Topic 22: star theater boring review great movie asleep beatles girl fell
Topic 23: star people review magic venue worth soleil great vega price
Topic 24: trick free people great stage seen boring worst time ticket
Topic 25: time better act star trick love beatles drink funny seen
Topic 26: magic act place music time boring really game beatles stage
Topic 27: boring better elvis minute song seat hour people asleep fell
Topic 28: elvis boring act night song pretty went funny place thought
Topic 29: vega food seat place trick line act audience star dance
Topic 30: better magic time people food performance funny service horrible ticket
Topic 31: act performance people trick vega water review venue good price
Topic 32: people elvis worth great david copperfield good magic minute price
Topic 33: vega water game funny acrobatics machine drink minute watch pinball
Topic 34: better act minute review pretty watch ticket music david cool
Topic 35: kid elvis act year drink think class music story blue
Topic 36: drink boring beatles david watch copperfield better asleep love fell
Topic 37: stage people kid disappointed love free entertaining star drink dance
Topic 38: performance pretty place water drink dance seen seat free little
Topic 39: better friend star performance people elvis disappointed service act staff
Topic 40: david disappointed copperfield review illusion boring acrobatics dancing vega water
Topic 41: pretty stage beatles cool experience david illusion song girl copperfield
Topic 42: watch disappointed free food seen story song went friend kid
Topic 43: place minute horrible recommend stage watch entertaining kid waste worst
Topic 44: watch love service funny worth elvis save entertaining venue customer
Topic 45: girl love half year asleep fell felt night performance music
Topic 46: music worth performer place girl review performance year elvis minute
Topic 47: entertaining girl review venue water blue high beatles food group
Topic 48: pretty review michael jackson love cool free high watch audience
Topic 49: half hour little kid watch think song save place reve

LSA Analysis: Top 50 important topics across high rating reviews

```
Topic 0: time great vega cirque seat love like really music beatles
Topic 1: machine pinball game place play arcade quarter hall fame cent
Topic 2: fountain bellagio vega water free night minute time watch beautiful
Topic 3: beatles cirque love music soleil fountain song machine acrobatics water
Topic 4: magic vega time family great kid funny trick love seen
Topic 5: great beatles love music arena song seat game fountain place
Topic 6: seat vega great best cirque ticket seen arena amazing price
Topic 7: machine great amazing water music really fountain stage performer pinball
Topic 8: time guy seen great amazing girl night bachelorette best second
Topic 9: game cirque really water place soleil like funny video great
Topic 10: time cirque good soleil fountain machine seen different ticket love
Topic 11: time amazing loved seat absolutely trick really entire second game
Topic 12: love place guy stage water beautiful audience amazing cirque girl
Topic 13: really vega good place amazing enjoyed love arena machine nice
Topic 14: ticket worth loved definitely cirque price soleil night friend free
Topic 15: water music arena beautiful like magic ticket people food line
Topic 16: loved vega water music seat love game ticket theater good
Topic 17: best free water music ticket amazing good seen guy magic
Topic 18: loved arena best cirque kid machine good concert seen food
Topic 19: good recommend water definitely jeff highly music entertaining family pretty
Topic 20: good amazing magic vega loved fountain music guy night game
Topic 21: magic trick definitely vega seat guy recommend arena water penn
Topic 22: place like awesome seat thing definitely water cool beatles free
Topic 23: place recommend highly music seen awesome year theater performance production
Topic 24: song watch good place beautiful water different night great worth
Topic 25: awesome definitely arena worth love free like watch song entertaining
Topic 26: like recommend seen highly definitely water great night funny love
Topic 27: worth definitely music theater night seat money audience seen seeing
Topic 28: people thing funny cool stage pretty audience highly loved recommend
Topic 29: watch night best like music beautiful minute stop recommend costume
Topic 30: free awesome night theater beautiful seat amazing funny song favorite
Topic 31: free theater beautiful entertaining drink nice time strip performance like
Topic 32: stage night awesome beautiful beatles seen family theater good experience
Topic 33: like audience funny awesome place performer time night best loved
Topic 34: beautiful watch seen place ticket entertaining funny favorite audience arena
Topic 35: watch theater funny people best come water year entertaining friend
Topic 36: funny beautiful best song experience went think thing going king
Topic 37: definitely favorite pretty cool stage drink best jeff ticket music
Topic 38: entertaining pinball family definitely night hall fame strip star minute
Topic 39: beautiful pinball definitely year good people jeff thing hall audience
Topic 40: pinball kid hall different fame pretty worth watch funny adult
Topic 41: minute strip better seen hour beautiful beatles funny sure dancing
Topic 42: funny year nice free stage pinball music performance seat favorite
Topic 43: performance definitely experience different seen enjoyed free performer year pretty
Topic 44: favorite thing strip performance bellagio star make enjoy minute funny
Topic 45: nice different thing performance theater minute going definitely guy funny
Topic 46: people kid audience favorite performance minute line free adult story
Topic 47: family seen friendly drink people absolutely fountain worth friend come
Topic 48: better think star arena theater free going performance little place
Topic 49: make sure entertaining kid guy trick story line penn fountain
```

Dataset Used:

1. business.json
2. review.json
3. user.json
4. vegas_cleaned.csv (NLP)
5. reviews_subset.csv (RS)
6. business_subset.csv (RS)
7. user_subset.csv (RS)

Jupyter Notebooks Used:

1. Data Preprocessing.ipynb
2. Natural Language Processing.ipynb
3. Recommender System.ipynb