ILS-Z534, Final Project: Executive Summary

Ahmad Al Marzook, Abhishek Babuji, Jay Kaiser, Jinsu Kim
December 12th, 2017

1 Task 1, Algorithm 1

1.1 Problem Definition

Using Information Retrieval across an index to suggest relevant businesses using weighting schemes.

1.2 Background

Information retrieval depends on taking a supplied query and searching an offline-built index for documents whose content best matches terms from the query.

1.3 Significance

This weighting system is derived from the belief that terms appearing in negative reviews must reflect negative qualities of a business, while terms appearing in positive reviews must reflect positive qualities. Therefore, by strongly-weighting terms that appear in positive reviews and punishing those of negative ones, documents retrieved from the inverted index created earlier should consist mostly of reviews with strong appearances of the positive terms and minimal appearances of the negative ones.

1.4 Algorithm Details and Experimental Design

Using Apache Lucene, a Java-based search engine library, an index of all reviews and tips from businesses in Toronto, ON from the Yelp dataset was created; each business was represented by a separate document in the index, and within that document, all reviews and tips created by users for that business were separately concatenated together and saved as an inverted-index to allow rapid and easy online lookup. For building a query, two separate weighting schemes were compared: ***vanilla*** and ***score-weighted***. In either instance, a random user is extracted from the Yelp dataset, and all of his reviews for any businesses are joined together into a single query. When building a *vanilla* query, these reviews are treated as equally important for lookup in the inverted index, independent of the total number of stars that users rated any businesses in his reviews. In comparison, a *score-weighted* query takes notice of the stars-rating for each review of that user and gives precedence to terms found in positively-rated reviews while punishing terms found in negatively-rated ones. Treating a three-star review as objectively-neutral (and thus not rewarding or punishing the appearance of terms found in such reviews in the index), additional- or fewer-starred reviews are given a positive or negative multiplier respectively. In this manner, a four-star and five-star review is given a multiplier of two and three respectively, while a two-star and one-star review is given a multiplier of negative two and negative three, also respectively.

1.5 Conclusion

Using Lucene, this system was completed, and the results of both weighting schemes are compared with Task 1's second algorithm later on in this paper, in comparison to the results from Algorithm 2.

2 Task 1, Algorithm 2

2.1 Problem Definition

Comparison of various similarity metrics for user-based and item-based Collaborative filtering method for the Yelp dataset.

2.2 Background

Collaborative filtering (CF) is a method of making predictions about a user's interest toward a particular item, given similar users or items. There are two techniques used to make these predictions. The first approach is user-based CF, where an implicit assumption is held that if a given user A is similar to a different user B in a certain respect; then there is good chance that user A is more similar to user B in other respects, and user A should hold more-similar opinions to user B than to that of some randomly chosen person. The second approach to collaborative filtering is the item-based collaborative filtering. Invented and used by Amazon.com in 1998, item-based CF considers the similarity between two items that a given single user may hold an opinion toward, instead of the similarities between two users, as is done in user-based CF.

2.3 Significance

There exists multiple ways to compute similarity between two users or two items. The user based similarity metrics also allow the specification of a neighborhood within which the top similarities will be considered. Task 1, Part 2 attempts to study the performances of these various similarity metrics for user and item-based collaborative filtering methods.

2.4 Algorithm Details and Experimental Design

Apache Mahout is used as an implementation framework to create user and item-based CF recommendation systems. Coded using Java, Apache Mahout provides many classes to implement various similarity metrics for both user- and item-based CF approaches. The original Yelp dataset represents *user_id* and *business_id* with unique strings. The input to Mahout

user_id	business_id	stars
0	0	5
0	4	5
0	51	2
0	107	1
0	115	2
0	134	5
0	136	1
0	178	4

Fig. 1. Input to Mahout

is 3 columns which are represented in the figure. The *user_id* and *business_id* columns are factorized in a way that each *user_id* and *business_id* is represented by a unique integer starting from 0. We are choosing the city of Champaign, IL which has 6713 unique users and 378 unique restaurant businesses.

2.5 Similarity Metrics

For user-based collaborative filtering, these are the similarity metrics which will be used for comparison.

`PearsonCorrelationSimilarity, LogLikelihoodSimilarity, TanimotoCoefficientSimilarity, EuclideanDistanceSimilarity, SpearmanCorrelationSimilarity`.

For item-based collaborative filtering, these are the following similarity metrics which will be used for comparison.

`PearsonCorrelationSimilarity, LogLikelihoodSimilarity, TanimotoCoefficientSimilarity, EuclideanDistanceSimilarity`. Since we're limited by the number of pages we're allowed to use, the math behind the various similarity metrics can be studied from their respective papers and will not be explained here.

2.6 Evaluation Metrics

2.6.1 MAE/Average Absolute Difference:

For each rating-prediction pair, their absolute error is calculated. After summing up these pairs and dividing them by the total number of rating-prediction pairs, Mean Absolute Error can be found. It is the most commonly used and can be interpret easily. [Taner Arsan, Efecan Köksal, Zeki Bozkuş]

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y_j}|$$

2.6.2 RMSE/Root Mean Squared Error:

This is a statistical accuracy metric that is slightly different from Mean Absolute Error. Once rating-prediction difference is calculated, its power of 2 is taken. After summing them up and dividing them by the total number of rating-prediction pairs and taking square root of it, Root Mean Square Error can be found. [Taner Arsan, Efecan Köksal, Zeki Bozkuş]

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y_j})^2}$$

Both the above measures almost represent the same thing. The reason for not using Precision and Recall is because, those are two evaluation metrics which are inherently dependent on the predicted recommendations for each user in the test set and the true item purchased. When it comes to deciding if your model is performing well, we need to look for the model that produces the lowest MAE or RMSE. A lower value of MAE and RMSE means that the predicted rating differs from the actual rating by that value, ideally you would want MAE and RMSE to be as small as possible.

2.7 Results

The following are the set of graphs that show the variation of performance with similarity metric and the neighborhood size.

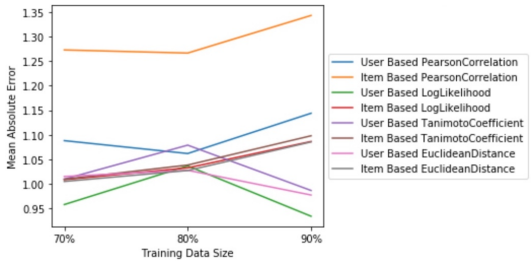

Fig. 2. USER BASED vs. ITEM BASED, MAE (NEIGHBOR = 10)

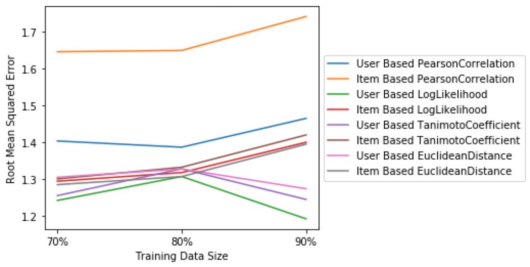

Fig. 3. USER BASED vs. ITEM BASED, RMSE (NEIGHBOR = 10)

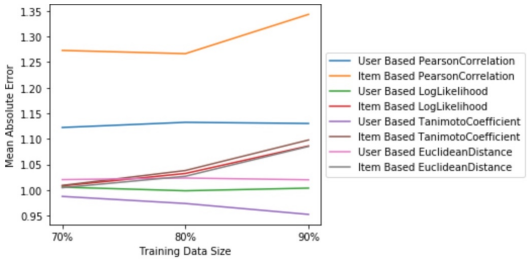

Fig. 4. USER BASED vs. ITEM BASED, MAE (NEIGHBOR = 25)

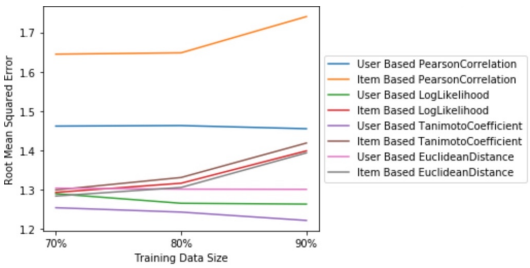

Fig. 5. USER BASED vs. ITEM BASED, RMSE (NEIGHBOR = 25)

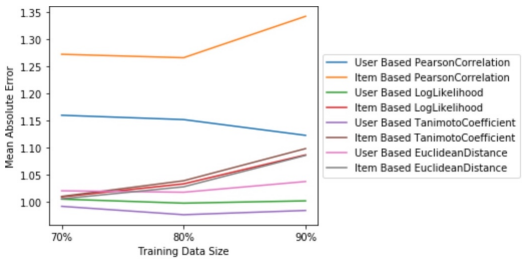

Fig. 6. USER BASED vs. ITEM BASED, MAE (NEIGHBOR = 50)

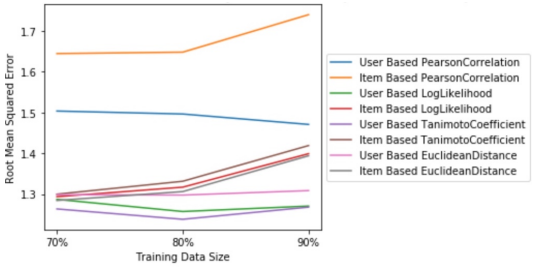

Fig. 7. USER BASED vs. ITEM BASED, RMSE (NEIGHBOR = 50)

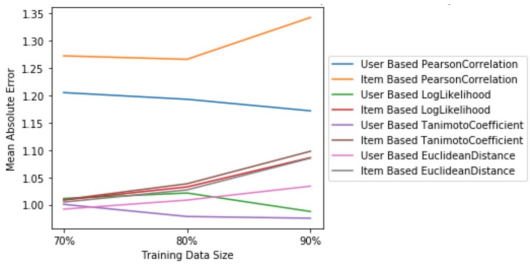

Fig. 8. USER BASED vs. ITEM BASED, MAE (NEIGHBOR = 100)

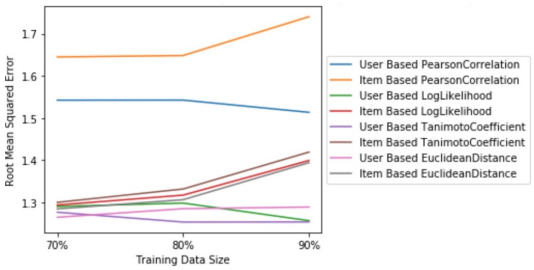

Fig. 9. USER BASED vs. ITEM BASED, RMSE (NEIGHBOR = 100)

2.8 Suggesting Recommendations

From the results above, it is clear that `UserBased TanimotoCoefficientSimilarity` with `100 NEIGHBORS` provides the best results. So for each *user_id*, 3 restaurant recommendations are printed using the above metric and neighborhood size. The recommendations are the restaurant *business_id* represented by unique integers the way it was taken in as Input.

3 Final Comparisons between approaches used in Task 1 - Algorithm 1 and Task 1 - Algorithm 2

To make a comparison (given the vast differences between Algorithm 1 and Algorithm 2 in practice) a random 70/30 split of training/testing data was created, and the training set was input as an index for Algorithm 1; queries are then created based on each user's remaining reviews in the training set. Accuracy is measured based on how many of a given user's own indexed reviews are returned in the top 100 recommended businesses. Because Algorithm 2's method for splitting datasets into training and testing sets is private, and because custom input of training and testing data is not allowed, the random splits for both algorithms are not identical. Results for Algorithm 1 are displayed below in Fig. 10., in comparison to the best and worst results from Algorithm 2. The best result from Algorithm 2 is obtained by the usage of `UserBased TanimotoCoefficientSimilarity` consistently providing low RMSE and MAE values (Refer to Fig. 4, Fig. 5, Fig. 6, Fig 7. and even Fig. 8 and 9, it is seen that `UserBased TanimotoCoefficientSimilarity` produces the least MAE and RMSE overall).

Given the single training/test split outlined above, statistics on accuracy, precision, and recall across both weighting schemes outlined in Algorithm 1 were found. Of the total number of users in the Champaign, IL Yelp dataset, only 1587 of 2986 users were successfully found in the indexed training set. The table below outlines the results found, where coarse precision or recall represent even a single user business retrieved within the 100 for the given user, and fine precision or recall represent only those outputs that contain all user businesses retrieved within the 100 for the given user.

Surprisingly, the "optimized" weighting scheme resulted in poorer results across the board. The reason for this is currently unknown. However, based on this measure of accuracy, the vanilla weighting scheme proved fairly accurate, especially given its crudeness in design. It is very difficult to attempt to compare the accuracy between Algorithm 1 and 2, given the nature of their different outputs. Algorithm 1 outputs suggested businesses given reviews, and Algorithm 2 outputs suggested star ratings for specific businesses given similarities to other businesses or other users. Accuracy for both have been found separately, but cross-comparing them in this research is similar to comparing apples to oranges.

	Vanilla Weighting	Scored Weighting
Accuracy among retrieved	0.6808973	0.539323
Accuracy across all	0.36188346	0.2866395
Coarse Precision	0.85444236	0.7038437
Fine Precision	0.43541273	0.34341526
Coarse Recall	0.45411924	0.37407905
Fine Recall	0.23141326	0.18251842

Fig. 10. Task 1 - Algorithm 1 Results

4 Task 2

4.1 Problem Definition

Task 2 centers among a simple question that can be easily derived given the nature of Yelp:

Can one see evidence of shifting economic trends (i.e. economic recessions and booms) by analyzing the information within the Yelp dataset?

4.2 Background

The philosophies underlying this research question derive from the impact the economy has on the average individual, *i.e.* the individuals who leave reviews on Yelp regarding their experiences at a given business. The impetus for this question is based on a real-world economic event, the 2008 Great Recession, whose impact led to drastic changes in spending habits across households, not only in America but in many other countries around the world.

4.3 Significance

Given the drastic consequences of this economic event, one could hypothesize spending habits (and hence reviews regarding spending habits) at the time would clearly align with overall economic trends. Moreover, given its diachronic dataset of both business reviews and new users, Yelp appears to be an ideal dataset for testing such changes of habits. It is with this in mind that the following hypotheses, on which our research has been based, have been designed.

4.4 Testing Various Hypothesis

4.4.1 Hypothesis 1: The total number of Yelp reviews will fluctuate with the economy.

This hypothesis comes most immediately to mind when considering the impact of the economy on consumer spending habits. Naturally, given a drastic downturn of the amount of wealth the majority of consumers find themselves possessing, the consideration of buying goods and services externally should also dramatically decline. By not going out to a business, a review for said business cannot be made. Therefore, there should almost be a one-to-one correspondence between the strength of the economy and the number of new reviews at a given time. To account for the possible impact the growing total number of Yelpers (i.e. Yelp reviewers) given Yelp's growing popularity, the total number of reviews for a given day were divided by the total number of Yelpers existent from the conception of Yelp to that day, and the resultant frequency was considered for analysis instead.

4.4.2 Hypothesis 2: The total frequency of positive (or negative) Yelp reviews will fluctuate inversely with the economy.

Given a sudden change of total capital for a consumer following an economic shift, his experience at an establishment will be shadowed by a comparison of the goods he is receiving to those he could receive for less elsewhere. Moreover, any poorer or unusual experiences at a business will be seen as a greater loss of time or money given a shortage of either of those than when neither play a determinant factor for the consumer. Therefore, negative reviews for a business should become easier to obtain during economic decline, and the same can be inversely said for positive reviews. It is under this concept that this hypothesis has been designed.

4.4.3 Hypothesis 3: The total frequency of Yelp reviews with mentions of words and phrases that reflect the economic health of reviewers will fluctuate with the economy.

This hypothesis considers the possibility that reviewers mention in their reviews the most important aspects of their experience with a given business or service. In times of economic hardship, the importance of cost of a good or service plays a greater role in a review than in would if money is more freely-available to spend; therefore, reviews during this time should contain greater mentions of economically-salient words and phrases. These may range from descriptors detailing a good or

service received at a business as either “cheap” in quality or “measly” in portion to “expensive” in price or “a good deal.” This hypothesis reflects the work done by Antenucci et al., (2014), who use mentions of unemployment or termination in Twitter data to try to similarly view economic trends.

4.4.4 Hypothesis 4: The total frequency of reviews to low-end businesses and high-end businesses will inversely fluctuate respective to the economy.

As consumers possess less capital with which to purchase goods and services, their source for such goods and services may shift down from businesses that offer higher qualities commodities and equally high prices to those that offer cheaper commodities at reflectively cheaper prices. For example, if an average middle-class consumer wished to purchase a hamburger, he would be presented with a number of options, ranging from higher-end establishments (e.g. Five Guys, Red Robin) to lower-end ones (e.g. McDonald’s, Burger King, Wendy’s). Given less income due to economic recession, a consumer should have a propensity to frequent the lower-end corner of the total spectrum of businesses, instead of their usual ones. The total quantity of review data from both lower-end and higher-end establishments should reflect this change of usual choice.

4.4.5 Hypothesis 5: The results found from Hypotheses 1 through 4 will undergo more extreme levels in cities and states harder hit by economic recession than in cities and states less-affected.

In cities or states where economic shifts either affect fewer consumers or affect consumers in a less-prominent manner, their consumption of goods and services should not change as drastically as will those who have undergone the brunt of an economic decline. Reversely, for those who have been most affected by a recession, consumer habits will change drastically from the norm to combat this. Therefore, the effects listed in the four hypotheses above should be seen more prominently and clearly for these individuals.

4.5 Methods and Experimental Design

Though these hypotheses provide a means to measure economic trends, no comparison can be provided without an economic baseline. Gross domestic product (GDP) and unemployment data are natural indicators of measuring this economic change over time. Overall GDP and the average GDP of the food and accommodation industry in particular US states were used for cross-comparison between differently-affected areas of the country (as is predicted to be beneficial given Hypothesis 5). GDP and unemployment data were retrieved from the Bureau of Economic Analysis (of the United States Department of Commerce) and the Bureau of Labor Statistics, respectively. When measuring Yelp frequencies, to account for Yelp’s monotonically increasing number of users over time, the total number of reviews for any given time were normalized by the total number of Yelpers who had joined Yelp from its inception to that time; these Yelpers must have left at least a single review for a particular state to be considered as a Yelper from that state.

4.6 Results

The Great Recession of 2008 affected GDP, both overall and in the food and accommodation industry, as well as unemployment rate of most states across the United States. GDP decreased sharply in the fourth quarter of 2008, showing slow recovery into the fourth quarter of 2009. GDP in the food and accommodation industry decreased somewhat in the first quarter of 2010, showing an even slower recovery in GDP than GDP overall. Viewing at government economic data, states that rely strongly on tourism (e.g. Nevada) were highly impacted by economic recession, showing an especially sharp decline in GDP in the food and accommodation industry.However, most of the review-frequency results that centered on the five hypotheses described above was found to not align with the overall economic trends described above, even in states that were most hard-hit. Across all five hypotheses, no evidence could be found that supported the notion that economic shifts impacted consumer trends on Yelp, despite expectations for the contrary. The graphs below show examples of each of the hypotheses when compared to the economy of Nevada, North Carolina, Ontario, and again Nevada over time, respective to each hypothesis; as can be inferred, none of the hypotheses graphs align strongly to the baseline graph. This does not mean, however, that from the Yelp dataset no inferences or points-of-interest can be made; for example, based solely on the total number of reviews found for most states, strong rapid peaks and nadirs in review frequencies can be easily seen, representing the change in consumer habits during the summer and winter, respectively. Despite the failure of the hypotheses, the appearance of this type of data suggests that Yelp does in fact provide consumer trends, just not given the hypotheses researched above.

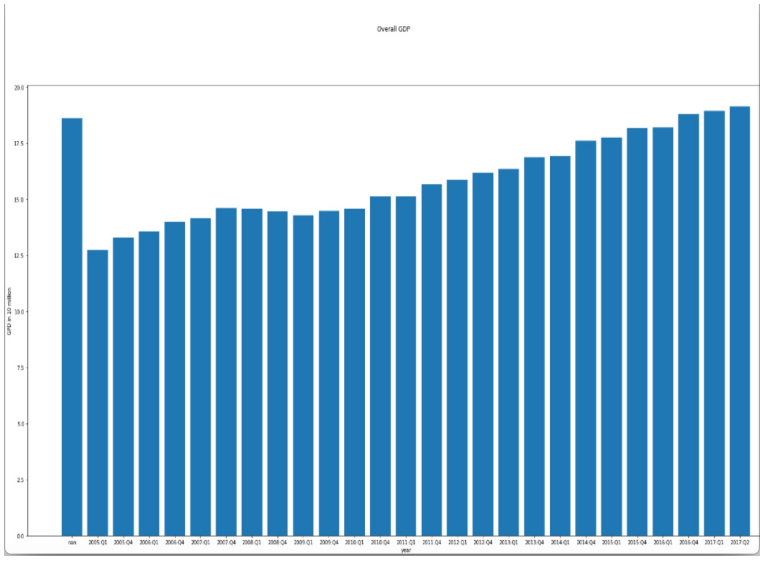

Fig. 11. United States overall GDP in millions over different quarters from 2001 to 2017

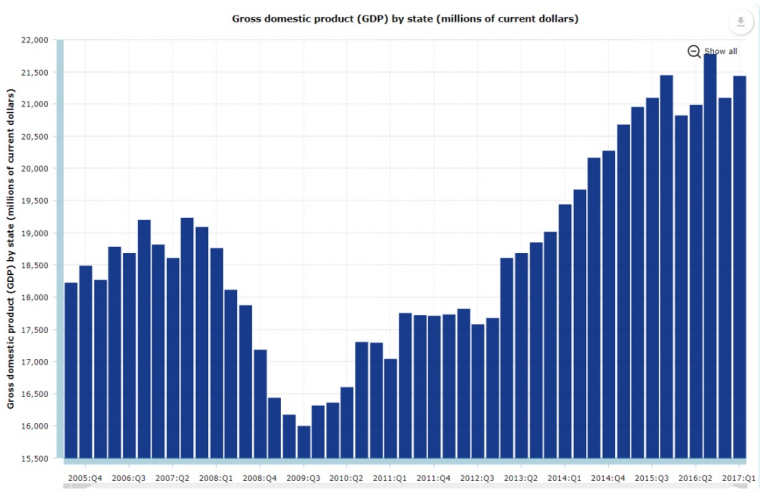

Fig. 12. Nevada overall GDP for the food and accommodation sector in millions over different quarters from 2005 to 2017

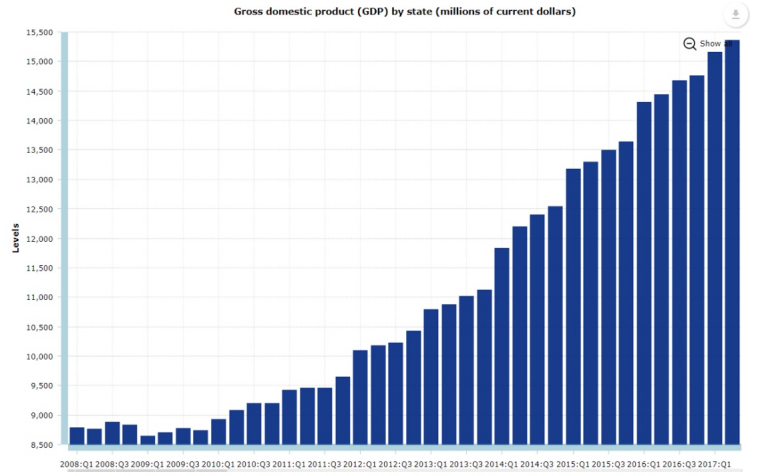

Fig. 13. North Carolina overall GDP for the food and accommodation sector in millions over different quarters from 2008 to 2017

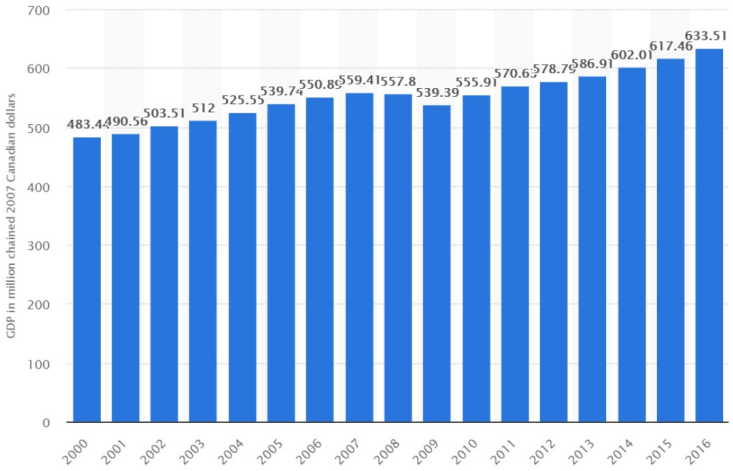

Fig. 14. Ontario overall GDP in millions from the years 2000 t0 2016

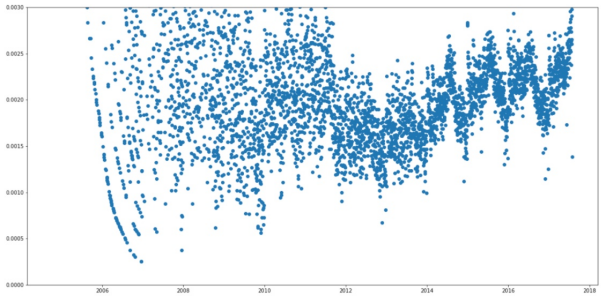

Fig. 15. Hypothesis 1: Total number of reviews in Nevada (normalized by number of Yelpers) from 2006 to 2018

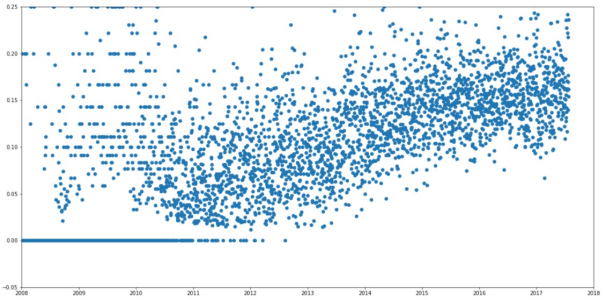

Fig. 16. Hypothesis 2: Frequency of reviews with a 1-star rating in North Carolina from 2008 to 2018

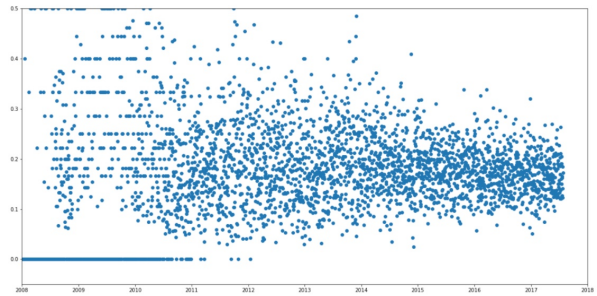

Fig. 17. Hypothesis 2: Frequency of reviews with a 2-star rating in North Carolina from 2008 to 2018

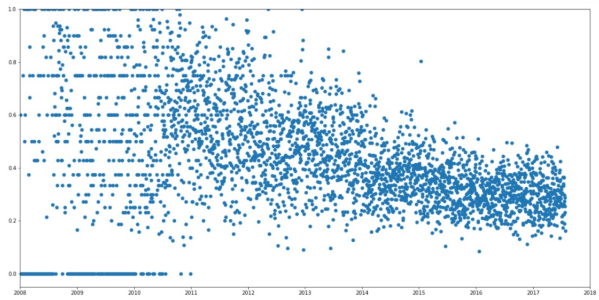

Fig. 18. Hypothesis 2: Frequency of reviews with a 3-star rating in North Carolina from 2008 to 2018

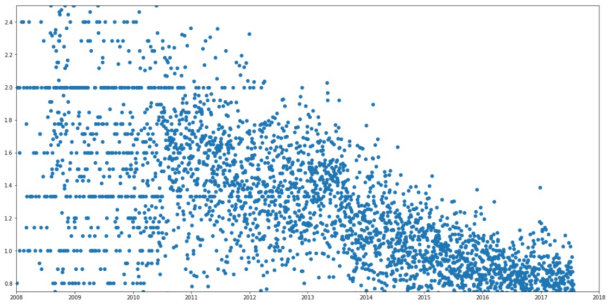

Fig. 19. Hypothesis 2: Frequency of reviews with a 4-star rating in North Carolina from 2008 to 2018

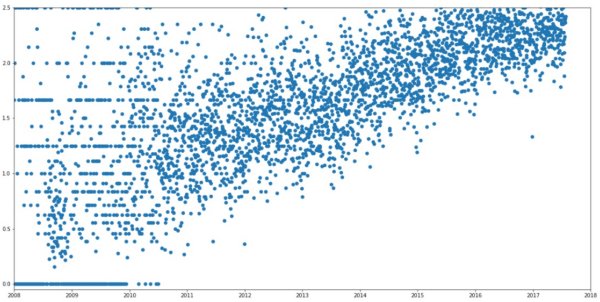

Fig. 20. Hypothesis 2: Frequency of reviews with a 5-star rating in North Carolina from 2008 to 2018

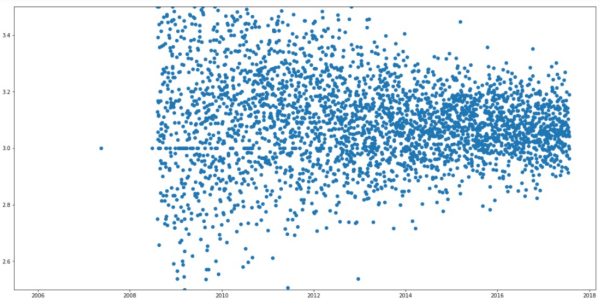

Fig. 21. Hypothesis 3: Frequency of reviews that mentions economically salient terms in Ontario from 2006 to 2018

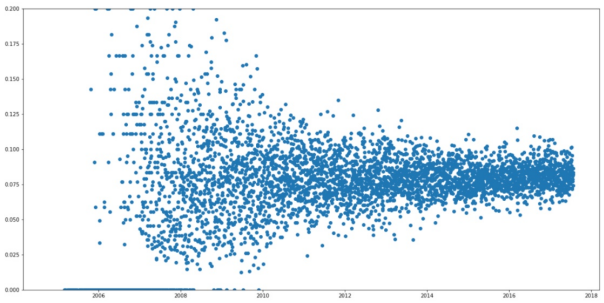

Fig. 22. Hypothesis 4: Frequency of reviews for low-end businesses in Nevada from 2006 to 2018

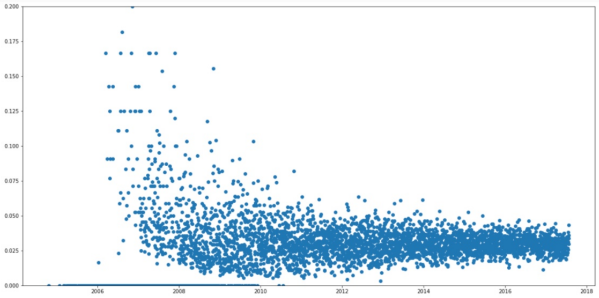

Fig. 23. Hypothesis 4: Frequency of reviews for high-end businesses in Nevada from 2006 to 2018

5 Conclusion and Limitations

Although the Yelp dataset utilized here contains almost eight million total reviews, the data is still wholly lacking in several key aspects. The most obvious limitation to this dataset is due to Yelp's late start as a popular review application; reviews only arrive in any significant measure starting around 2008. Because the Great Recession only happened during this time as well, there is no baseline present to allow us to compare the number of Yelp reviews before and after the recession. This means that although conclusions could possibly be made from the data regarding any economic shifts in the years following the Great Recession, the most easily-analyzable years for such a study as this are unavailable.

Another limiting factor to this dataset is the few number of reviews overall, as well as the lack of diversity to the locations from which these reviews originated. To properly compare cities strongly and weakly hit by the recession, certain cities (e.g. Oklahoma City, OK and Detroit, MI respectively) would provide more obvious insights to the effects of the recession on consumer trends; however, these cities (among any others hoped to have been studied here) are missing from the dataset. Additionally, many states and cities do not have enough reviews for even general trends to be measured. Instead, levelled stratification occurs in the data points, yielding no useful correlations to be extracted.

6 Future Work

In future iterations of this project, more focus should be given to review and user data from the Yelp dataset, as well as outside sources of that too could provide economic trending data. In particular, trends for individual groups of users can provide the bulk of analysis, as opposed to the overall counts of reviews that were utilized here. To account for the data sparsity encountered during this project, additional Yelp data can be scraped, not only adding more data to work with overall, but also providing a means to compare data from specific cities of interest; by targeting cities known to be hit the hardest during economic shifts, a better analysis can be made.

Additionally, other factors recorded by Yelp in their dataset can also be considered as factors for measuring economic shifts; these include service quality, average waiting time, parking space availability and price, all of which can provide better indicators of the success of a business or its popularity than star-ratings in reviews alone. Because most of Yelp's bulk data only begins starting in 2008, the most important year of study for this research, other datasets from this time, like Twitter, may prove necessary to gather enough data to possibly provide a comparison of popularity of businesses across consumers. Finally, future work on this project could include measuring the overall sentiment of each review, allowing more obvious means of comparing overall consumer-satisfaction from before and after a major economic shift.

7 References

1. Antenucci et al. (2014). Using Social Media to Measure Labor Market Flows. National Bureau of Economic Research. Working Paper.

2. Arias et al. (2016). Metro business cycles. Journal of Urban Economics. Volume 94, 2016, 90-108.

3. Bureau of Economic Analysis U.S. Department of Commerce. Retrieved December 1, 2017, from https://www.bea.gov/.

4. United States Department of Labor Bureau of Labor Statistics. Retrieved December 1, 2017, from https://www.bls.gov/home.htm/.

5. Comparison of Collaborative Filtering Algorithms with various similarity measures for Movie Recommendation, International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.6, No.3, June 2016 by Taner Arsan, Efecan Köksal, Zeki Bozkuş; Department of Computer Engineering, Kadir Has University, Istanbul, Turkey.