ILS - Z534

Information Retrieval: Theory and Practice

Project Report

Yelp Dataset Challenge

Fall 2014

Indiana University Bloomington

Ghanshyam Malu
Puneet Loya
Suhas G Ramakrishna
Suprith Chandrashekharachar

December 14, 2014

Abstract

Yelp produces crowd-sourced reviews about local businesses. It also registers information pertaining to businesses such as hours of operation, their categories, location and more. Yelp’s dataset can be mined to obtain useful knowledge about the businesses ultimately helping both the businesses to improve their services and customers to make an informed decision.

In this project, we are attempting to predict and refine already existing categories for a business based on user reviews and tip information. Additionally, we are addressing a problem to predict a set of best positives and negatives of a business. The analysis and prediction is done using information retrieval techniques.

Table of Contents

1 Introduction

In the first task we improve the category metadata of the businesses from the Yelp Dataset and evaluate the existing categories of the businesses.
For example:
Business: “Prairie Land Service Center”
The categories of a business from yelp are: ["Auto Repair", "Automotive", "Towing"].
Here, the categories are few and does not signify the different services provided by the business.

In the second task, our **research question** is to bring out the positives and negatives of a business.
From the results of the above method, the system can offer the customers to:

- Choose a best among the lot in a business domain.
- Avoid the need for reading long reviews.
- Save time and prevent subjective bias.
- Conclude what to expect from a business in a glance.

2 Task: 1 - Predict Categories of Businesses

2.1 Proposed Solution:

Following steps were executed to achieve the objective of improving the category metadata of the businesses:

i. Import the Yelp dataset into MongoDB
ii. Create Different types of Analyzers for indexing
   a. **XAnalyzer**: Analyzer with removal of stop words.
   b. **BAnalyzer**: Analyzer with an option for multi gram indexing along with stop words removal
iii. Index “business”, “review” and “tip” information of all the businesses using Lucene for unigram, bi-gram and trigram
iv. Get the top words for all unigrams, bi-grams and trigram using the algorithm
v. Get the master list of all the categories available in Yelp
vi. Compare the obtained top words with the master list to propose the categories for the each business.

2.2 Algorithm

The algorithm followed for this task is described below:

a. For each Analyzer
   - Create a collection of words from business (name), review (text) and tip (text).

b. For each word in the collection get the score.
   - score ← α * score$_{name}$ + β * score$_{review}$ + γ * score$_{tip}$
   - score$_{name}$ ← TF$_{(business)}$ • (IDF$_{(business)}$)
   - score$_{review}$ ← TF$_{(review)}$ • (IDF$_{(review)}$ + IDF$_{(ap89corpus)}$)
   - score$_{tip}$ ← TF$_{(tip)}$ • (IDF$_{(tip)}$)

c. Top words ← words sorted based on score

d. Top 10 of these words are matched with existing categories master list to predict the new categories.

α,β and γ are the weights given to fields business name, reviews and tips respectively.

There are two IDF (Inverse Document Frequency) parameters considered:

1. IDF of the particular field.
2. *IDF of the term in ap89Corpus dataset*. This IDF is introduced to reduce the noise, i.e., overused terms across two datasets must be ignored as they may not contribute significantly to categorize the business. This IDF is applied only while calculating the term score from the review field of the dataset.

2.3 Evaluation

Due to the unavailability of a standard evaluation tool which can evaluate the efficiency and performance of our algorithm, we have chosen Precision and Recall metrics to evaluate the results from our algorithm.

The formulas used to calculate the metrics are described below.

$$\boldsymbol{Precision = \frac{Number\ of\ correctly\ predicted\ categories}{Number\ of\ predicted\ categories}}$$

$$\boldsymbol{Recall = \frac{Number\ of\ correctly\ predicted\ categories}{Number\ of\ \text{given}\ categories\ \text{in dataset for a business}}}$$

2.4 Experiments

As the number of categories defined across different businesses vary, businesses with categories more than 8 were indexed to measure the precision and recall. Since the dataset is large, experiments have been conducted over a subset of data to avoid excess usage of computational resources. The experiment results are depicted below.

The graph in Figure 1 shows the difference between including IDF from an external dataset and with its exclusion. The external dataset we used is APCORPUS89. Out of the 51 businesses compared, 7 businesses using external IDF shows negative results. All other businesses show positive results where precision is either equal or better by utilizing IDF from external dataset.

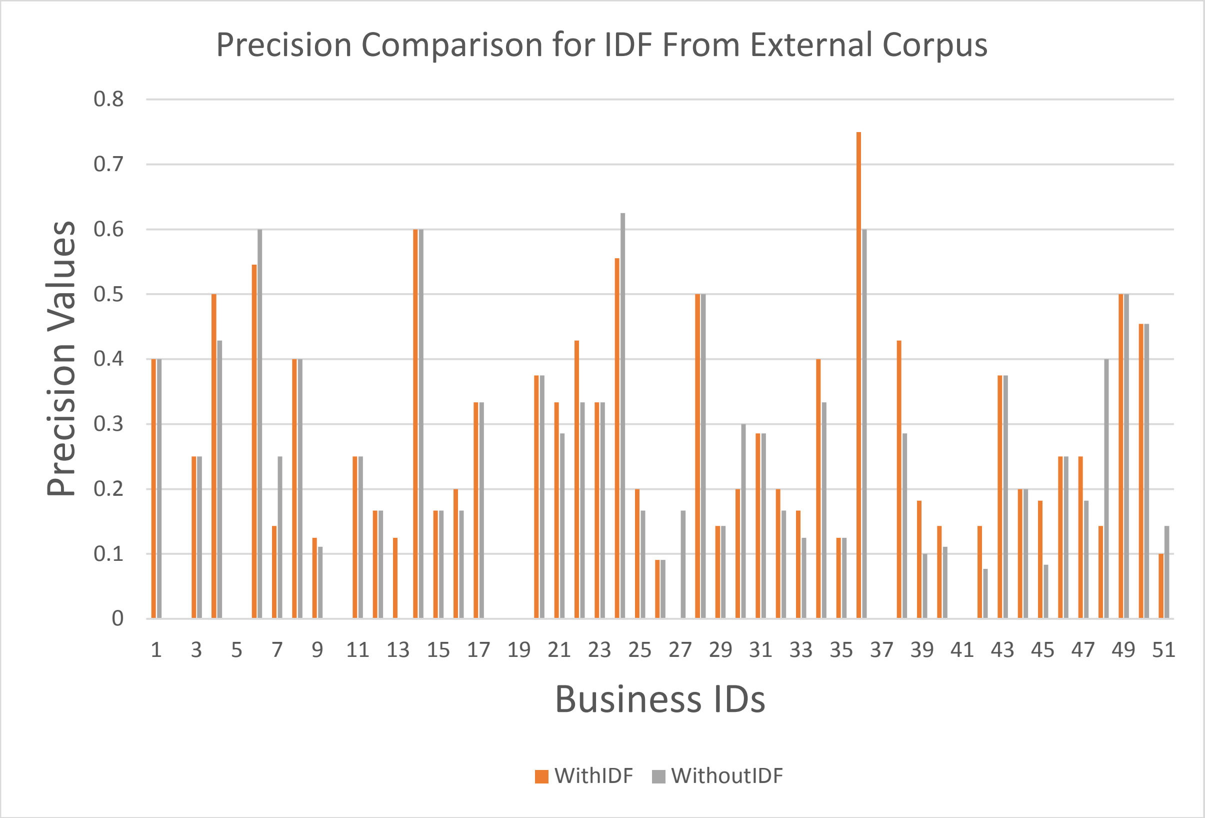

Figure 1

The graph in Figure 2 indicates the precision and recall of category predictions. As the categories are sparse for businesses the recall is quite high and precision is low.

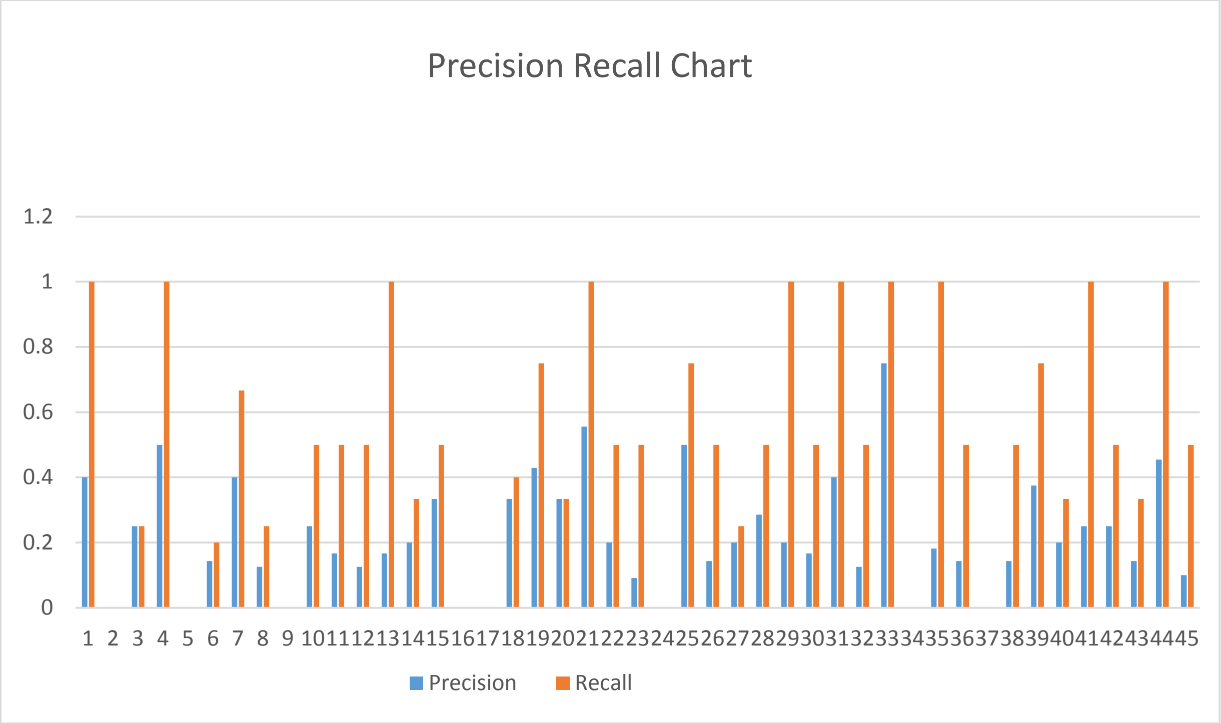

Figure 2

2.5 Tuning Field Weights (α, β, γ):

This section disseminates information about the experiments conducted to determine the best set of weights to enhance the algorithm's performance – recall and precision. Due to resource constraints, experiments were performed on a set of 20 businesses. The graphs in Figure 3 and Figure 4 draws comparisons across two tests with different set of values for different fields namely: Review, Tip, and Business Name.

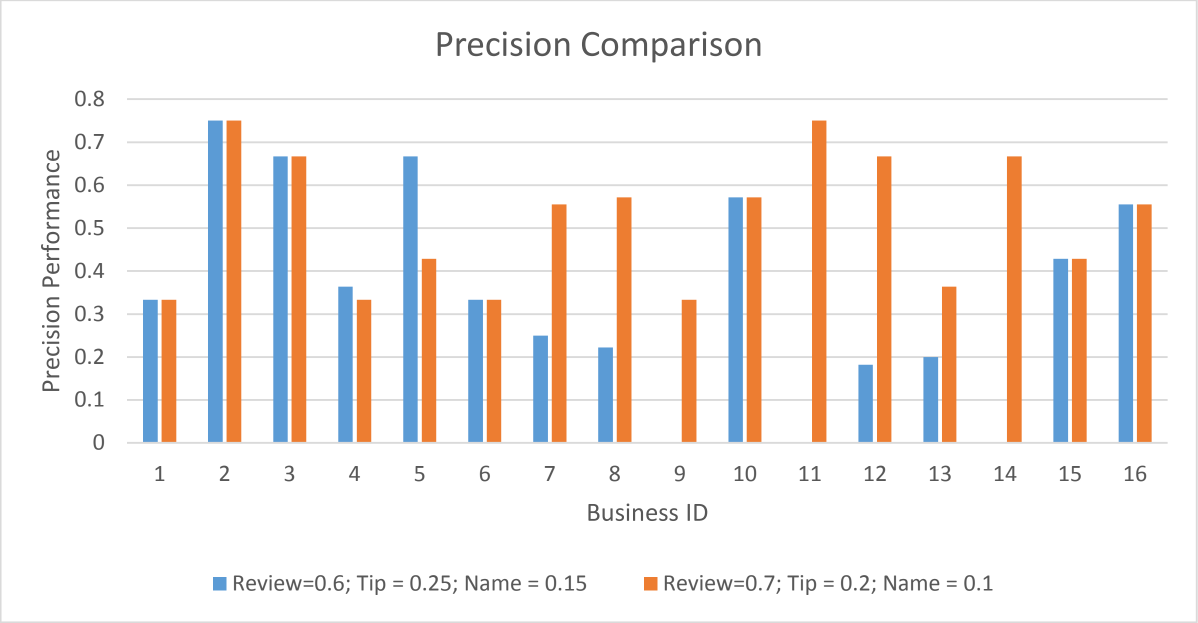

Figure 3

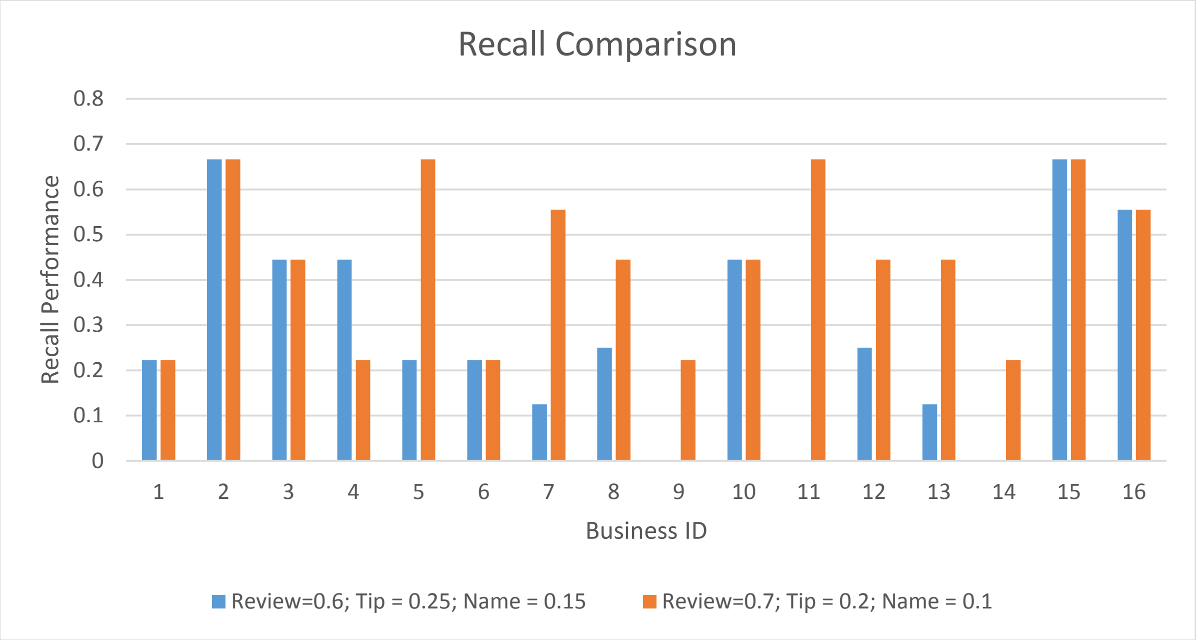

Figure 4

We have experimented with wide range of values for review, tip and name. As the weight of review filed was increased to 0.8, the precision has been observed to go down. Since all the other set of values didn't impact the algorithm's performance by a good amount, Review=0.7; Tip = 0.2; Name = 0.1 was chosen as optimal set of values.

2.6 Limitations

- Noise elimination: Typos in reviews may be neglected and words that are not contributing to the category must be well filtered.
- No semantic association, hence the precision may vary.

2.7 Future Work

- We can improve the precision by experimenting the IDF from a larger neutral data set.
- Provide category recommendation.

3 Predicting top 5 positives and negatives of the business

In this task, we attempt to bring out the positives and negatives of a business.

3.1 Algorithm

The algorithm followed for this approach is described below:

1. Initiate StanfordNLP package for sentiment analysis and noun extraction
2. For each business
   - Top trigrams obtained from Task 1
   - For each trigram : $P_{score} = 0$, $N_{score} = 0$
     1. Get all the sentences from the reviews which contain that trigram along with their rating [0-5].
     2. For each sentence
        (1) Calculate the sentiment of the sentence using StanfordNLP.
        (2) Get Positive$_{score}$ or Negative$_{score}$
        (3) $P_{score} = P_{score}$ + Positive$_{score}$
        (4) $N_{score} = N_{score}$ + Negative$_{score}$
     3. Normalize S_{core} and N_{score} based on the count of positive and negative sentences
     4. If P_{score} or N_{score} outweigh each other and passes the threshold **θ**
        (1) Get nouns of trigram using StanfordNLP
        (2) Add it to best/worst list.

3.2 Sentiment, Positive$_{score}$ and Negative$_{score}$

High level flow for obtaining the Sentiment, Positive$_{score}$ and Negative$_{score}$ is shown below.

- For each sentence obtain Sentiment value using StanfordNLP
  1. Sentiment is assigned as per the below scale
     a. 0 – Extremely Negative
     b. 1 – Fairly Negative
     c. 2 – Neutral
     d. 3 – Fairly positive
     e. 4 – Extremely Positive
  2. Sentences with Neutral sentiment are ignored
  3. Adjust the Sentiment to CorrectedSentiment based on whether the sentence is positive or negative sentence.
  4. If Positive Sentiment
     - Positive$_{score}$ = CorrectedSentiment * rating[0 - 5]
  5. Else
     - Negative$_{score}$ = CorrectedSentiment * rating[0 - 5]

3.3 Evaluation

- The threshold θ is kept very high, around 4 (on 5 scale) to make sure that the answers are genuine.

- It means that the average rating for that positive/negative word is 4
- Therefore there is chance that No positives or Negatives appearing for a particular business.
- The evaluation is done manually for selected business by reading reviews by non-developers (Peers) and then compare to analyze scores with the output of the program.
- The observations made by the non-developers are taken as truth table.
- Precision and Recall are calculated using the below formalue:

$$\boldsymbol{Precision} = \frac{\boldsymbol{Number\ of\ correctly\ predicted\ postives\ and\ negatives}}{\boldsymbol{Number\ of\ predicted\ postives\ and\ negatives}}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{Number\ of\ correctly\ predicted\ postives\ and\ negatives}}{\boldsymbol{Number\ of}\ \textbf{positives and negatives}}$$

3.4 Experiments

To get the best positives and negatives of a business, the problem was first manually solved by two peers who did not have any prior knowledge about the results obtained by program. The comparison of the results of peers and our results show huge difference. The difference in the methodology is we extract top trigrams and evaluate their positive or negative nature based on their usage in the sentences and approach used by peers is pure judgement.

Example 1:

Business Name:	John & Kathy's Smoke Shop	
	Manual	**Output from Program**
Positive	Friendly; Courteous; Good service; Location	
Negative	Charge extra for credit card	charge; Credit card; post office experience

Precision = (2/3) = 0.66
Recall = (2/5) = 0.4

Example 2:

Business Name:	Flux	
	Manual	**Output from Program**
Positive	Variety of items; Unique Item; Recycled Products; Good cards and jewellery	gift shop; minding; jewellery; bit;
Negative	Expensive; out of town	treasure cavern; baby slipper; emporium

Precision = (1/7) = 0.14
Recall = (1/6) = 0.16

In this example we get a very low precision and recall.

The cause observed for low precision and recall in our approach is lack of tuned algorithm for better prediction. Also, manual predictions from more peers must be obtained and average out the results to create a more refined truth table.

3.5 Limitations

- Sparseness: If the reviews are sparse for a business, there might be positives or negatives just based on these small number of reviews.
    - Example: Only one review with 5 stars is present in a business will give positive words only from that review
- There were very less negative reviews in the observed set. Most of the negative reviews have less number of stars. The accuracy for negatives is always low compared to positives.
- The threshold is a subject of discussion. Fixing on threshold is yet to be perfected.
- Sometimes the words in the trigrams are not meaningful. Relying completely on nouns from trigrams may not be the best way.
    - Example : “Town” , “Day”

3.6 Improvements

- Alternatives for extracting meaningful words and reducing noise from the sentences need to be explored.
- Our results comply with the StanfordNLP package for sentiment analysis. The package is not 100% accurate. Other packages have to be explored.

4 Conclusion

We have mined the yelp dataset and predicted categories for each business using TF-IDF approach and predicted the top 5 positive and negative reviews. Addressing both the research questions in one project was our goal. During the project implementation, we have had to deal with multiple challenges. Prime challenges have been to retrieve words that can predict a category from informal review texts, implementing custom analysers to handle bi grams and tri-grams with some text pre-processing, and figuring out external libraries for natural language processing as well. Additionally, our algorithm for predicting categories had to be tuned to enhance the performance by giving different weights to reviews, tips and business names.

StanfordNLP made our life easier in crafting a solution for the second research problem, which involved state of the art techniques like sentiment analysis on text inputs. Due to time and computational resource constraints we managed to experiment predicting categories, positive and negative reviews for few businesses and results have been presented in a concise manner.

5 Project Website

https://github.iu.edu/suhgulur/Yelp_team5

6 References

The Stanford NLP APIs are used from:

http://nlp.stanford.edu/software/corenlp.shtml

7 Credits

Shravan Kumar and Jagadish Madagundi for helping with the unbiased ground truth contribution.

8 Other Submitted Documents

- Ground Truth for top positives and negatives of a business: Ground Truth.xlsx
- Output log for top positives and negatives of a business: outputlog