Augmenting choices

By extracting subtopics from Yelp user reviews and identifying non-intuitive business categoreies

Shubhadeep T Bhattacharya
Dept. of Computer Science
Johns Hopkins University
Baltimore, USA
Sbhatt14@jhu.edu

Abstract—

In this paper, the intent is to identify subcategories of businesses and services by studying user reviews on Yelp. Further, this study was extended to identify categories that are non-intuitive. These subcategories can help businesses boost their revenue, while users can quickly locate businesses and services based on popular subcategories. Identifying non-intuitive sub categories will enhance the profile of a particular neighborhood as these are some of the most unique businesses that are listed on Yelp. I used the Yelp academic open dataset that has 4.1 Million reviews. To find the latent subcategories by studying the reviews, I employed the Natural Language Took Kit, to compute n-gram and their frequency of occurrences. For identifying the non-intuitive categories 1.1 Million business attributes that are tagged by the users were combined with the business categories for three cities. Overall a lot of interesting insights were found this could definitely help the users and businesses alike.

Keywords

Yelp, Subcategories, tagged attributes, standalone subcategories, n-gram features

1. INTRODUCTION

Yelps rating on businesses have impact on their revenues. On an average it has been observed that each star rating adds an overall 5% - 9% increase in revenue [1]. It is thus in the best interest of the business to be listed on Yelp. How could the business leverage this to drive their revenues? This analysis is carried out to see what the most trending topics or terms of a particular city are. The idea is that by including the topics as subcategories, users can have more varieties within the trending topics.

Some of the business categories that are present in Yelp includes – Restaurants, Shopping, and Food. Within these broad categories lies the subcategories like American (New), Chinese and Spanish within the Restaurant.

Attributes of the business makes them unique. These are the features that they offer. This could be outdoor seating, casual dining and Wi-Fi availability for restaurant or insurance accepted and wheelchair friendly for hospital and surgery clinics. The attributes are tagged by the users and can take value of True and False. Like restaurants accepting credit card True.

Identifying non intuitive categories, i.e. category and a true attribute can help Yelp list business in an interesting way. For example Korean bar offering karaoke. Users can also select attributes to narrow their searches within the subcategory. Using tokenization of adjacent elements of user reviews and calculating the frequency, the most tagged attributes could be derived. Topic modelling of the reviews by using LDA enables us to get features which are common to most of the business within a given city. Identifying these features as a standalone subcategory would better enable the users to search for the most popular restaurant.

BACKGROUND

Sentiment analysis of the Yelp dataset that focused on predicting star reviews and predicting business category using clustering has been done. The highest star rating restaurant were clustered and the most important features were identified as business categories [2]. The analysis was comprehensive but the features identified for a 1 and 3 star restaurants were 60% identical.

Other approaches that were carried out to factor models for discrete data included were dimensionality reduction techniques. Popular among them is the Latent Semantic analysis (LSA) and Latent Semantic Indexing. With the advent of Probabilistic Latent Semantic Indexing (PLSI) which models each words in the document as sample of mixture model [3]. For small document its performance suffers from over fitting

Extensive work has been carried out for studying online review site. [Chevalier and Mayzlin 2003] discussed that the review feature such as word count in book review website are important to potential customer and not the star ratings, thus applying K Nearest Neighbors of features on 5 star restaurants on Yelp will result in lot of good business to be left out thereby excluding subcategories.

Analysis on Groupon coupons on Yelp business reviews has shown that a lot of coupons are redeemed just before the expiration date hurting business rating [4].

One approach using LDA and Expectation Maximization algorithm that converges to most likely parameters to discover latent topics has been carried out by Huang J, Improving Restaurants [6]. This approach has been successful in finding out subcategories for made up categories that are defined manually. The limitation of this approach being derived subcategories are limited to categories that are identified manually.

I use NLTK n gram analysis and frequency counts on user reviews to find the trending topics, this could be included as categories. Also, the attributes that are true were included with the categories to find the non-intuitive categories.

For topic modelling Gensim is used. Gensim incorporates Latent Dircihlet Allocation (LDA) factor model to approach unsupervised learning of topic in a given text. Similarity retrieval with large corpora is efficient with Gensim and thus it was used on the user reviews to identify popular subcategories.

3. METHOD

A. Dataset

The research is done using the Yelp academic dataset [5]. This dataset contains business, user reviews, user information, user checkin and tip in each separate JSON objects.
The business JSON object contains information about the business id, name, city, attributes, categories and type. The user review JSON object contains information on the text review, stars etc.
The checkin JSON object contains information on the time, business id and type. The tip JSON object contains text of the tip, date, business and user id.

B. Tools

The research was carried using Python script, using the Natural Language Toolkit and Gensim Library for topic modelling. Visualization was done using Wordcloud.

C. N-Gram Features with Preprocessing.

N-gram features use concatenated words in order to capture key features between words and phrases.

Unigram of user review with English stop word and analyzing the frequency for a given city with different category. This was extended to find the bigram with the corresponding frequency and the trigram and the corresponding frequency. A lot of interesting insight has been observed while analyzing the higher n-grams.
Business category are combine with attributes that are true to form tokens. Tri-gram and quad-gram along with their frequency may help identify non-intuitive categories.

D. Word2Vec.

Word2Vec allows words representation learning using continuous bag of word model. Words that are similar linguistically are closely related to each other and are closer to each other in the vector space.

For the key feature identified with the English stop words using n-gram preprocessing, retrieve the feature word using, model [‘feature’]. The query of the ‘feature’ words will result in retrieval of words that are close in vector space and are based on cosine similarity. These feature words can be identified as subcategories. Many of these feature word can be a standalone category/subcategory.

E. Topic Modelling

For topic modelling using the Yelp user reviews for restaurants in Calgary resulted in the following topics at relevance matric $\lambda = 0.79$ of 50% of the tokens.

Topic	Frequency
Cocktail	4.2%
Pasta	3.9%
Dishes	3.8%

TABLE 1. Distribution of Topic in user reviews in Calgary

Since Pasta and Cocktail have a high frequency and are amongst the most relevant terms including them in category would be an option.

4. Results

A. Categories/Subcategories

The first case study was done for the city of Edinburgh, Scotland. Yelp has placeholders for categories which is common for cities all across the world. Some of the categories/ subcategories presently listed on Yelp Edinburgh based on all business in the city.

Categories	**Frequency**
Tea	2%
Coffee	1.8%
Shopping	1.7%
Pubs	1.6%
Laundromat	< 0.0001 %
Electricians	< 0.0001 %

TABLE 2. Distribution of Category on Yelp Edinburg

B. Category/ subacategory by user review

Analyzing user reviews popular unigram that can be part of Yelp subcategories.

Unigram most popular	**Frequency**
Offer	0.005%
Haggis	0.007%
Tapas	0.006%
Curry	0.005%

TABLE 3. Distribution of Category on Yelp Edinburg

Bigram popular	**Trigram popular**
Staff Friendly	Devonshire clotted cream
Well worth	Old famous restaurants
Reasonably Priced	White Hart Inn
Live Music	Within Walking Distance
Afternoon Tea	Sticky Toffee pudding

TABLE 4. Alternative subcategory suggestions based on Bigram and Trigram analysis of user reviews.

Trigram analysis of user reviews resulted in identifying food, culture and businesses that are unique to the place.

Trigram popular	**Cities**
Pittsburg Primanti Brothers	Pittsburgh
Genral Tao Chicken	Montreal
chocolate pistachio croissant	Montreal
Corned Beed Cheese	Pittsburgh

TABLE 5. Popular trigrams identified as subategories

C Infering subcategories by Word2Vec.

The words that are closer in vector space to the preexisting Yelp categories/subcategories can be utilized as an independent categories/subcategories.

Existing categories	**Possible subcategories based on Cosine similarity**
Food [category]	Menu, chicken, pork, burger, dining, meal, sandwich, fries, sushi
Sushi [subcategory]	Zen, rolls, sashimi, teriyaki, tuna
Burger [subcategories]	Fries, sandwich, delicious, satisfying

TABLE 6. Inferring subcategories by cosine similarity

In Edingburg naan, curry and tandoor has cosine similarity with Indian. Pubs serving curry dishes are relagated to end of the search results and are often missing. A more varied subcategories within Indian cusine can mitigate this.

D Discovering non intuitive corelations among categories.

The attributes that are frequently tagged in Pittsburg, USA.

Fig 1. Frequently tagged Attributes.

Attributes present in business that are popular were combined with business categories, n-gram analysis to find non-intuitive categories these can be identified as unique business and help improve profile of a neighborhood. These would be a whole new way to list businesses.

('street', 'caters', 'goodforkids', 'lunch')
('surrounding', 'greenery.', bar)
('interior', 'classy', touch 'light', 'industrial')
(‘clubs’, goodfordancing, 'hvac’)

TABLE 7. Non-intuitive categories

5. Conclusion

Using Wod2Vec I was able to infer new subcategories with cosine similarity to the preexisting subcategories. It turned out that providing a variety of subcategories would help users locate a business more effectively, Curry vs. Indian. N-gram frequency analysis of user reviews provided insights on how Yelp could recommend features that are unique to a place Pittsburgh Primanti Brothers in Pittsburgh, USA and General Tao Chicken in Montreal, Canada.

Features that businesses supports were studied to infer correlation among non-intuitive categories example, street caters good for kid’s lunch. Listing business with non-intuitive categories would be an innovative way to draw more users.

Topic modelling of user reviews have indicated that the users care about cocktails, services, tasty, beautiful, reasonable, price and take out. Categories could be better modelled around these terms.

6. Future Work

Online LDA Algorithm can be employed on the entire user reviews to obtain features that user care about. Cosine similarity of these features could be analyzed to infer new categories.

Training an n-gram language model and estimating the probability of a sentence [7] [8], in our case non-intuitive categories would help identify the unique business that should be listed on Yelp.

a.

7. Acknowledgment

I offer my sincerest gratitude to my instructors at Johns Hopkins, who have supported me throughout the research. Thanks to Dr. Ian McCulloh and Dr. John Piorkowski for being excellent instructors who have motivated me in completing the research.

References

[1] Blanding, Michael. "The Yelp Factor: Are Consumer Reviews Good for Business?." *Harvard School of Business* (2011).

[2] Hung, Kevin and Qui, Henry 2014 "UCSD Yelp Dataset Challenge" http://kevin11h.github.io/YelpDatasetChallengeDataScienceAndMachineLearningUCSD/#show-naive-bayes-math.

[3] Hoffman, M., Bach, F.R. and Blei, D.M., 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp.856-864).

[4] Inman, J.J. and McAlister, L., 1994. Do coupon expiration dates affect consumer behavior?. *Journal of Marketing Research*, pp.423-428.

[5] Sourcs: https://www.yelp.com/dataset_challenge

[6] Huang, J., Rogers, S. and Joo, E., 2014. Improving restaurants by extracting subtopics from yelp reviews. *iConference 2014 (Social Media Expo)*

[7] http://masatohagiwara.net/training-an-n-gram-language-model-and-estimating-sentence-probability.html

[8] Stolcke, A., 2002, September. SRILM-an extensible language modeling toolkit.In *Interspeech* (Vol.2002,p.2002