Inferring Business Similarity from Topic Modeling

[Latent Dirichlet Allocation and Jaccard Similarity applied to Yelp reviews]

Alex Yang*
University of California San Diego
9500 Gilman Dr.
La Jolla, CA 92122
aly006@ucsd.edu

Carlo Provinciali
University of California, San Diego
9500 Gilman Dr.
La Jolla, CA 92122
cprovinc@ucsd.edu

ABSTRACT

In this paper, we apply the discovery of subtopics in Yelp reviews towards determining the level of similarity between Yelp businesses. Specifically, we apply an online Latent Dirichlet Allocation (LDA) algorithm to extract the latent subtopics within all reviews for a given business. We then aggregate the most frequently occurring subtopics per business, and utilize them to calculate the Jaccard Similarity between the topics of other businesses. Utilizing this combination of latent topics and similarity measurement allows us to uncover both subtle and obvious relationships between the reviews of two establishments. In addition, we describe other discoveries including the variance in topics found by rating distribution, as well other insights gained by an inspection of latent topics in review text.

Categories and Subject Descriptors

H.4 [**Software Engineering**]: Data Science; D.2.8 [**Data Mining and Predictive Analysis**]: Topic Modeling—*Latent Dirichlet Allocation*

General Terms

Natural Language Processing

Keywords

Topic Modeling, Latent Dirichlet Allocation, Jaccard Similarity

1. INTRODUCTION

With over 70 million reviews and 142 million monthly unique visitors, Yelp is a treasure trove of information for small businesses. The purpose of this paper is to take the reviews submitted by these users and process them in an effort to determine the high-level latent characteristics of a business, and to use these characteristics as a similarity measure between businesses. This has applications ranging from business recommendation to determining a particular business' competitors.

The core idea is that a given review submitted by a user is likely to contain latent characteristics of a business itself within the text. For example, from the sample review below, we can make the reasonable prediction that this business is a restuarant which serves Mexican cuisine:

> Decent burritos at late hours. The portions are fairly large and tasty. I ordered the Surf and Turf for $7.40. There is ample outside seating area and the service is very fast! The main draw for me is that it is open late.[1]

We may also observe that important features such as the service and hours are mentioned. We can apply Topic Modeling algorithms such as Latent Dirichlet Allocation (LDA) to learn and extract these characteristics (topics) from a corpus of reviews. While the latent topics found in one review may not necessarily be representative of the business, by aggregating the latent topics over multiple reviews, we can observe distinct characteristics of a business that emerge. For a restaurant, these characteristics might include cuisine, service, hours, ambience, and/or affordability.

Once we can represent the characteristics of a business by uncovering its most highly correlated latent topics, we can utilize the characteristics to estimate the business similarity with other businesses. The review above is taken from Vallartas, a popular Mexican restaurant in San Diego. In this paper, we will show how we can use the characteristics of restaurants such as Vallartas, to determine other restaurants a user may enjoy, or what other businesses might be a likely competitor.

2. EXPLORATORY ANALYSIS

The dataset we analyze in this paper is the Yelp Academic Dataset[2]. The dataset itself contains business, review, and user information for the 250 most popular businesses across 30 different universities in the United States.

*Website: aly006.github.io

[1]http://www.yelp.com/biz/vallarta-express-mexican-eatery-san-diego

[2]https://www.yelp.com/academic_dataset

The dataset is contained in a single JSON file and is divided in three categories: businesses, users, and reviews. For the purpose of this task, we focus on the business and review objects within the data. Each review includes the following fields:

```
{
  'type': 'review',
  'business_id': (the identifier of the business),
  'user_id': (the identifier of the authoring user),
  'stars': (star rating, integer 1-5),
  'text': (review text),
  'date': (date, formatted like '2011-04-19'),
  'votes': {
    'useful': (count of useful votes),
    'funny': (count of funny votes),
    'cool': (count of cool votes)
  }
}
```

Similarly, the business data includes basic information about each business including name, business id, category, location, geographical coordinates, average star rating, etc.

We notice that although the dataset contains on average roughly 24.5 reviews per business, the median of reviews is only 4. Therefore, we can conclude that the majority of businesses will have a very small number or reviews, and the distribution of reviews per business is skewed to the right. In addition, the categories 'food' and 'restaurant' appear among 49 out of the 50 most popular businesses.

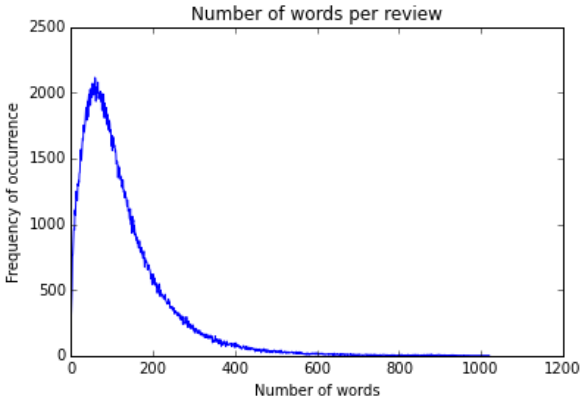

Figure 1: Frequency of word distribution per review

Figure 1 refers to the distribution of reviews according to their length in words. We observe that the average length of each review is 127 words while the median is 99, which suggests that review length distribution is not as skewed as reviews per business. We can reasonably estimate that reviews in the fiftieth percentile have an approximate corpus of 400 words in total. This suggests that models that are based off latent topics in review text for this dataset may be generally more informative in comparison to models that are based off latent dimensions in ratings or other features alone.

Figure 2 shows the comparison of frequency of words per review for the two categories of businesses. The red line represents businesses in the Restaurants category, and the blue line represents all non-Restaurant businesses. From this plot we observe that the review corpus is dominated by reviews of food related businesses, and that the distribution of words per review is extremely similar across all businesses.

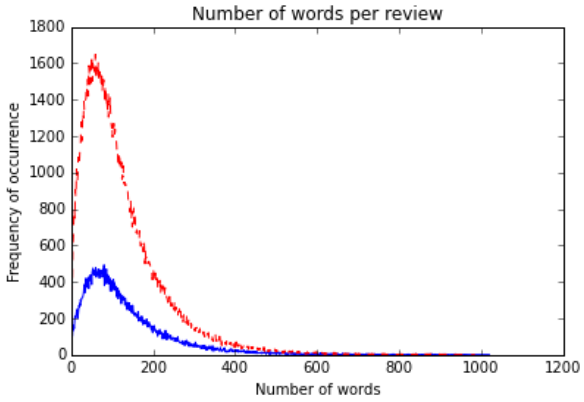

Figure 2: Frequency of word distribution for food-related vs non food-related businesses

3. RELATED WORK

Thanks to the incentive offered by the Yelp Dataset Challenge[3], a number of academic papers have explored the potential of topic modeling on the Yelp dataset. The vast majority of literature seems focused on utilizing this type of analysis as a way to infer users' sentiment that could be helpful in predicting rating criteria.

For instance, the work of Huang et al.[3] highlights how Latent Dirichlet Allocation[1] can effectively be used to describe the latent topics in restaurants reviews. In particular, they discover that topics related to service, value, take-out, and decor appear more often in user reviews, which can be helpful to identify the users' evaluation criteria. Furthermore, they attempt to associate a rating for each hidden topic by taking the average over the reviews that displayed the corresponding topic. The metric is meant to aid business owners in identifying which aspects of their business are particularly appreciated by the customers and which ones need improvement. This analysis provides good evidence of how online LDA[2] - which allows for faster and more memory efficient training compared to the traditional LDA model - can be particularly valuable for latent topics analysis tasks with a significantly large number of documents.

The Hidden Factor Model designed by McAuley et al.[4] utilizes a combination of latent dimensions in rating with latent topics in review text to gain more accurate insights from rating. This approach provides a sensible improvement from the existing models by fusing hidden topics in product review text as a parameter in standard recommendation tasks, leading to greater accuracy in prediction. This is particularly beneficial to in solving the so called "cold start problem" of rating predictions based off latent rating dimensions, as it is difficult to generate predictions for users that

[3]http://www.yelp.com/dataset_challenge

have given only few ratings. Intuitively, the topics discussed in a single review can give much more information regarding a particular user or business compared to the rating score alone.

Although there are several examples that attempt to utilize review text to predict rating and customers' sentiment, there doesn't seem to be much evidence of efforts to predict the similarity between businesses based off topic modeling of review text. This is the predictive task which we describe in the following section.

4. PREDICTIVE TASK

The predictive task we perform upon this dataset is primarily one of topic modeling. Specifically, we aim to extract latent topics from a corpus of Yelp reviews using Latent Dirichlet Allocation[1], as described by Blei et al. To achieve this, we use a combination of several tools, including python libraries such as Gensim[4], NLTK[5], and PyMongo[6], to perform LDA, process text, and persist data, respectively. The general implementation work-flow is described in the following subsection.

4.1 Implementation

4.1.1 Data Storage

The Yelp Academic Dataset is given in the format of a single JSON file, with each object separated by a newline delimiter. The file contains a mixture of business objects, review objects, and user objects, whose structure is described above in section 2. We use the *PyMongo* library to process the business and review objects in the dataset, and persist them into separate MongoDB collections. We filter all objects by the 'Restaurant' category, as we concern ourselves only with the subset of data containing Restaurant reviews.

We then determine the relevant features to be stored in each collection. The review text is naturally chosen, as it is a necessary component for the topic modeling task we aim to perform. Further processing of the review text is described in the section below. Business and review identifiers are stored for easier reference and mapping between collections. Other notable features we store include the number of reviews for a given business, to be used for observing differences in accuracy in predicting topics in businesses with a wide distribution of review counts. The business location is also persisted, to use geographical location to gain insights about nearby competitors or possible business recommendations.

4.1.2 Preprocessing Review Text

We apply several techniques to clean the review text, in an effort to simplify analysis. Looping through each review object in the collection obtained above, we convert the entire review to lower-case while simultaneously stripping punctuation. We then tokenize the review into words, while removing stopwords using the set of stopwords provided by NLTK. Finally, we use a Lemmatizer to find the lemma of the remaining words, in an effort to further strengthen words that belong to a single lemma.

[4]https://radimrehurek.com/gensim/
[5]http://www.nltk.org/
[6]http://api.mongodb.org/python/current/

We use the Gensim library to create a dictionary mapping the cleaned word tokens to identifiers, and save this dictionary locally. During this process, words with the lowest frequency are discarded. The resulting dictionary is also used to convert the cleaned text into a Bag-Of-Words corpora, also stored locally in Matrix Market[7] format. At this point, the major components for performing the topic modeling are in place.

4.1.3 Latent Dirichlet Allocation

The LDA model is a probabilistic generative model that is useful to automatically discover topics from a text corpus. It is based off the assumption that documents are composed by a mixture of topics (which are chosen according to a Dirichlet distribution over a fixed set of topics) and these topics generate given words with certain probabilities (based on the topic's multinomial distribution). More specifically, each document is seen as a bag of words, where each word is generated at random by the topics that define the document.

Suppose we want to discover a set of K topics. Since both the words contained in K and the topic composition of each document is unknown, we can utilize an algorithm such as Gibbs sampling to discover to learn the topic representation of each document and the words associated to each topic.

More specifically, each document is processed and each word within it is assigned to one of K topics. We then compute the probability of a topic appearing in a given document and the probability of a word appearing in a given topic; this procedure is repeated for each word and topic in the data and use it to adjust the words in each topic. After a several repetitions of this sampling, the assignments to each topic will eventually converge. Due to the size of the review corpus, and the relatively large number of iterations typically required for a standard LDA model to reach convergence, training the model can be very computationally expensive. In our analysis, we set K=50 and utilized the online LDA from the Gensim library, which provided a much faster and less computationally intensive convergence condition, making the training time feasible for our purposes with our resources.

In general, LDA has been shown to be more effective than other latent text models such as Latent Semantic Indexing.[1] The latter uses Singular Value decomposition to select words with high weight and combine the ones with high cosine similarity. Although this process is much faster - since it only needs one iteration to be trained - it is usually less precise and during evaluation and testing we found that LSI displayed more overlapping between topics than desirable.

4.1.4 Calculating Similarity

With the LDA model in place, we can predict the latent subtopics of a given corpus of review text. Utilizing the collections created above, we find all reviews of a given business and use the LDA model to find the distribution of topics for each review belonging to the business. We then find the most prevalent subtopics found among all reviews of a business, and compare the results to the aggregate topics of another business, via Jaccard Similarity, shown below.

[7]http://math.nist.gov/MatrixMarket/formats.html

Jaccard Similarity 1. The Jaccard coefficient measures similarity between finite sample sets A and B, and is defined as the size of the intersection divided by the size of the union of the sets A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

A Jaccard Similarity score above a certain threshold implies a high correlation between the latent characteristics of two businesses. The results of these findings are discussed in the **Results** section below.

4.2 Evaluation

We apply our online LDA algorithm for a range of topic numbers T ∈ [10, 100]. After cleaning the review text, only the top 10,000 occurring words by frequency are considered. We ultimately find that T = 50 performs well in properly determining and separating latent subtopics. For topic numbers in the low range T ∈ [0,25], vocabulary from separate topics becomes over-generalized and grouped into single topics. For topic numbers in a high range T ∈ [75,100], topics obtained by LDA become too specific, and seemingly similar vocabulary will become distributed among multiple individual topics.

We also evaluate our LDA model against other Topic Modeling techniques such as Latent Semantic Indexing (LSI). We evaluate the accuracy of our latent topic Jaccard Similarity measure against a simple baseline measurement of Jaccard Similarity between the categories of businesses. In both cases, we expect the results obtained by our model to be more descriptive and accurate.

5. RESULTS

A sampling of latent subtopics and their breakdown derived from training the LDA model are shown in **Table 2** (limited to 8 topics to preserve space). Some interesting results we observe include the presence of latent topics that represent the sentiment of a review, such as T0 and T6. Other topics such as T10 represent service. Unsurprisingly, we observe that several topics such as T2, T8, and T11 represent cuisine. A representation of all the topic breakdowns as more general "subtopics" is shown in **Table 1**.

We can use the model obtained from LDA to predict latent topics in previously unseen review text. **Table 3** shows the results of our LDA model upon a Vallartas review taken directly from Yelp. The percentage breakdown of topics is shown, sorted by greatest correlation. We observe that the model has successfully discerned that the review is a Mexican Restaurant and that it has a positive sentiment. We also notice a distinct drop in accuracy of topics in the lower percentages of topic correlation.

We obtain the most prevalent topics for a given business by applying our LDA model to the corpus of reviews belonging to a business. First, we derive several topic distributions of the business as shown for one review in **Table 3**, and store the most prevalent topics (those that exceed a certain threshold percentage P). Jaccard Similarity is then applied among the most prevalent topics between two businesses.

Table 1: General Subtopic Representations

#	Topic
0	Positive (Overall)
1	Location
2	Indian Cuisine
3	Negative (Taste)
4	Family
5	Location 2
6	Positive (Overall)
7	Lunch
8	Mexican Cuisine
9	Nightlife
10	Service
11	Burgers and Fries
12	Review
13	Unhealthiness
14	Service 2
15	Rating
16	Greek Cuisine
17	Seafood
18	Italian Cuisine
19	Breakfast
20	Location 3
21	Value
22	Pizza
23	Ambience
24	Thai Cuisine
25	Service 3
26	Return Visit
27	Parking
28	Quality
29	Meat
30	Mexican Cuisine 2
31	Vietnamese Cuisine
32	Late Hours
33	Service 4
34	Service 5
35	Hot Dogs
36	Sandwiches
37	Fondue
38	Cafe
39	Japanese Cuisine
40	Dessert
41	College
42	American Cuisine
43	Bar
44	Negative (Experience)
45	Wait Time
46	Positive (Experience)
47	???
48	Cake
49	Space

Table 2: Sampling of Latent Subtopics Derived From LDA

T	Breakdown
0	0.061*good + 0.058*like + 0.053*really + 0.033*pretty + 0.030*place + 0.024*dont + 0.017*think + 0.017*much + 0.017*little + 0.016*im
2	0.047*indian + 0.039*chicken + 0.037*buffet + 0.033*food + 0.027*curry + 0.022*naan + 0.022*dish + 0.020*rice + 0.016*spicy + 0.014*mango
6	0.071*great + 0.050*place + 0.046*food + 0.036*love + 0.031*good + 0.023*delicious + 0.023*awesome + 0.022*service + 0.020*priced + 0.017*quick
8	0.125*taco + 0.058*chip + 0.051*fish + 0.039*margarita + 0.032*salsa + 0.023*mexican + 0.023*tapa + 0.016*sangria + 0.013*enchilada + 0.012*tortilla
9	0.060*drink + 0.042*place + 0.040*bar + 0.026*fun + 0.023*friend + 0.023*night + 0.021*great + 0.017*music + 0.016*go + 0.015*cool
10	0.031*u + 0.029*table + 0.026*minute + 0.019*order + 0.017*wait + 0.015*time + 0.015*food + 0.012*came + 0.010*got + 0.010*took
11	0.162*burger + 0.119*fry + 0.023*onion + 0.019*cheese + 0.015*good + 0.014*potato + 0.014*ring + 0.013*sweet + 0.012*bun + 0.011*bacon
19	0.073*breakfast + 0.067*egg + 0.051*brunch + 0.031*toast + 0.030*pancake + 0.026*waffle + 0.021*french + 0.020*diner + 0.018*sunday + 0.018*morning

Table 3: Latent Topics in Vallartas Review

Text	This is my go-to Mexican place. First off, it has a drive through so my lazy butt doesn't have to change out of pajamas if I'm having a craving for a Cali Burrito. Fast service, good quality and good prices. Their tacos are also really good! I got the carne aside taco combo plate (with rice and beans) and loved it. Also if you order the chips and quac they give you a huge box full of hot crispy tortilla chips smothered in fresh guacamole. Big enough to share between at least 3-4 people, I can never finish it all because the portion is so big.
24.9%	0.054*food + 0.044*place + 0.032*good + 0.027*price + 0.017*get + 0.017*go + 0.015*pretty + 0.014*better + 0.012*cheap + 0.012*decent
18.69%	0.112*burrito + 0.052*mexican + 0.041*chipotle + 0.033*salsa + 0.026*nacho + 0.022*tortilla + 0.021*la + 0.018*quesadilla + 0.015*guacamole + 0.015*bean
11.57%	0.052*chicken + 0.037*sauce + 0.026*rice + 0.023*meat + 0.022*pork + 0.021*fried + 0.017*beef + 0.017*bbq + 0.016*spicy + 0.013*korean
10.48%	0.061*good + 0.058*like + 0.053*really + 0.033*pretty + 0.030*place + 0.024*dont + 0.017*think + 0.017*much + 0.017*little + 0.016*im
9.56%	0.125*taco + 0.058*chip + 0.051*fish + 0.039*margarita + 0.032*salsa + 0.023*mexican + 0.023*tapa + 0.016*sangria + 0.013*enchilada + 0.012*to rtilla
7.74%	0.042*time + 0.023*first + 0.022*back + 0.018*year + 0.017*went + 0.016*im + 0.015*last + 0.014*still + 0.013*one + 0.012*since
3.72%	0.071*great + 0.050*place + 0.046*food + 0.036*love + 0.031*good + 0.023*delicious + 0.023*awesome + 0.022*service + 0.020*priced + 0.017*quick
2.92%	0.023*dont + 0.023*like + 0.020*food + 0.019*make + 0.016*place + 0.015*want + 0.015*eat + 0.014*go + 0.014*vegetarian + 0.014*know
2.63%	0.088*wing + 0.064*falafel + 0.061*wrap + 0.057*crepe + 0.024*tip + 0.021*pita + 0.021*chicken + 0.020*sauce + 0.019*buffalo + 0.018*tax
2.44%	0.067*chocolate + 0.065*cake + 0.050*cupcake + 0.027*pastry + 0.023*sweet + 0.020*car + 0.018*frosting + 0.016*cream + 0.016*red + 0.014*mini
2.21%	0.015*paper + 0.011*napkin + 0.011*ketchup + 0.009*table + 0.008*water + 0.008*tray + 0.008*heat + 0.007*bus + 0.006*menu + 0.006*lamp
1.82%	0.051*ann + 0.050*arbor + 0.035*garden + 0.032*cookie + 0.026*daughter + 0.020*island + 0.019*palo + 0.019*alto + 0.017*fake + 0.017*ucsd

Table 4: Restaurants Most Similar to La Burrita

Restaurant	Category	Stars
Gordo Tacqueria	Mexican, Restaurants	3.5
Fat Slice Pizza	Pizza, Restaurants	3.5
Pancho's	Mexican, Restaurants	3.0
Top Dog	Fast Food, Restaurants	4.0
Mario's La Fiesta	Mexican, Restaurants	3.5

Table 4 shows the restaurants near University Of California, Berkeley with the maximum Jaccard Similarity in comparison with the restaurant La Burrita. We observe that the restaurants with the highest similarity unsurprisingly include those which serve the same cuisine. However, we can see an additional dimension in results such as restaurants that may be open during similar hours, have similar levels of service, or have similar prices. We evaluated our similarity measure for percentages for $P \in [0,15]$, and found a threshold of $P = 6$ to yield the most descriptive results. Threshold percentages that were too low yielded several restaurants that were only loosely similar, and percentages that were too high yielded restaurants that were incredibly similar. Indeed, an analysis of the restaurants with the maximum Jaccard Similarity with Top Dog in Berkeley, CA for high levels of P yielded only other Top Dog locations within the same city.

We compare the results of our model against a simple baseline of a typical Jaccard Similarity performed on the categories of businesses. This baseline typically yielded results in line with what we would expect; namely, given a restaurant La Burrita with the categories Mexican and Restaurant, it would output several Mexican Restaurant businesses (often with a maximum similarity of 1). While these are reasonable results, they lack the depth that our model provides by looking at the corpus of review text of a business. For example, while the baseline Jaccard Similarity model based upon categories is effective at determining restaurants with matching cuisine, it cannot differentiate by other significant factors including pricing, service, wait time, and ambience.

Finally, we observe differences in prevalent subtopics based upon review rating. As can be shown in **Table 1**, latent topics with both distinct positive and negative sentiment were discovered by our model. For reviews of a given business, a higher rating correlated with the appearance of latent topics with positive sentiment, while a lower rating correlated with the appearance of latent topics with a negative sentiment. A latent topic with positive sentiment typically contained several positively associated words, such as T0 in **Table 2**; the opposite was true for topics with negative sentiment. While a 1-star and 5-star rating of the same restaurant often ended up with similar cuisine and temporal subtopics, there were consistent differences in the sentiment topics, which included healthiness, taste, and overall experience. The significance of these results is that our LDA model can also determine sentiment in an unseen user review, in addition to topics. By aggregating the most prevalent topics across all reviews of a business, we can get a general opinion on how users evaluate the restaurant in aspects as subtle as the healthiness of the food or service and wait time.

6. CONCLUSIONS

With application of the LDA topic modeling algorithm, we discover latent topics within Yelp reviews. We use these topics to observe that review text often reveals information such as location, ambience, service, and cuisine. Furthermore, we notice that positive and negative sentiments are represented in latent topics, and successfully apply this information to observe the differences between reviews of high and low ratings. Furthermore, we aggregate the results of these latent subtopic distributions within reviews to determine the most prevalent topics among reviews for a given business. Finally, we use these prevalent topics to represent the characteristics of a business, and apply a Jaccard Similarity measure upon these characteristics. We find that our similarity measure is able to capture subtle similarities between businesses that are difficult to capture by more primitive models, and compared our results to a baseline that calculated similarity based on business categories. For future work, we hope that these insights will be useful to restaurants and users, and perhaps improve business recommendation or rating prediction.

7. REFERENCES

[1] A. Y. N. David M. Blei and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, January 2003.

[2] M. Hoffman and D. M. Blei. Online learning for latent dirichlet allocation. 2010.

[3] S. R. James Huang and E. Joo. Improving restaurants by extracting subtopics from yelp reviews. July 2013.

[4] J. McAuley and J. Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013.

Figure 3: Prevalent Topics Across All "Top Dog" Restaurant reviews