Topic-Based Tag Matching of Implicitly Similar Businesses

Kenta Takatsu, Caroline Chang

October 22, 2017

1 Research Context

Yelp contains 1240 tags that associate each business with discrete categories such as *Restaurants* or *Shopping*. Each business gets one or more tags to help the users identify the business category and these tags range from very broad topics to specific subtopics. These tags are usually generated by Yelp users or Yelp's manual data curation teams; however, this category tagging is not optimal in terms of the user experience – particularly on interpreting meaningful contents – for various reasons. First, some tags contain ambiguity that allows multiple interpretations across the users such as *Real Estate Services* and *Real Estate Agents*. Second, the categorical representation of business might not best describe what the business is actually known for. For example, business with *Pizza* tag might be known for their dessert menu, not pizza. Finally, the users might be more interested in subtle information about the services that can bridge across the different categories, for example *good portion for price* and *friendly staff*. In order to extract these implicit features, we need to investigate the raw review texts which contains direct 'feedbacks' from the users. Particularly, we focus on topic modeling algorithms such as Latent Dirichlet allocation (LDA) to cluster business based on the similarity of such implicit reviews topics. We can apply this study to improve user experiences by 1) defining less ambiguous tags by using the distribution of high dimensional topic embeddings, 2) understanding the most relevant and representative features about the business, 3) continuously generating new tags from subtle topics that exist across different business. In a long term, better business tags can improve the search engine, the filtering, and the recommendation system of Yelp.

2 Research Objective

The objective of this study is to discover, extract and quantify the performance of the high-denominational embeddings of business features. These features are later used to measure the vector distance between 2 businesses, which determines the similarity of them in a high-dimensional space. This way, we can construct a new metric that substitutes the existing 'business tag' which limit the similarity of business within the scope of predefined discrete bins. Also we can verify the similarity value as a linear combination or multiple topics, for example business1 and business2 share 80% of TopicA and 20% of TopicB. We are going to use LDA topic model and its empirical distribution as a potential candidate for the embedding algorithm; however, we already discovered that each review – even about the same business – contains large variance in its topic distribution and hence we need to generalize the embedding even more. Currently, we are exploring an mean distribution of topics over certain time-intervals to establish the baseline. Later, we will apply methods such as Word2Vec and Stacked Denoising Autoencoder to generate embedded features.In order to validate the performance of our embedding, we will make an assumption that business tags in Yelp data capture the accurate representation of businesses. With this assumption, we are able to conduct experiments from both supervised and unsupervised approaches.

3 Research Road Map

As previously mentioned, this research can be split into supervised experiments and unsupervised experiments. For both experiments wen are required to construct an LDA topic model across various business categories and apply statistical methods to validate our work.

3.1 Experiment Design

We have already constructed both NMF and LDA models using *Pizza* and *Chinese* as a pilot run. We need to carefully define the scope of topic models since our main objective is to extract subtle features in the review, and therefore, we are not interested in filtering out *Chinese* business from *Restaurants*. Rather, we are more interested in implicit differences between *Food Court Chinese* and *Authentic Chinese*, both of which do not exist as predefined Yelp tags. After constructing latent dirichlet allocation for each review, we will process this distribution by applying stacked denoising autoencoder to further generalize the topic model. This embedding will be used for both supervised and unsupervised learning method.

For the supervised method, we will train the binary classifiers to predict if 2 businesses share the same tags based on our embedding. Hence, our training set contains both embedding and tags for the businesses and our test set only contains the embedding. Since the previous studies by Yelp engineering showed promising results with an ensemble method such as Random Forests, we will use the same model as a baseline and see if our embedding can improve the tag matching. During this phase, we need to investigate false positives and false negatives, as they might indicate that the original tags were not perfect representation of the business. For this reason, it is crucial to use precision and recall to measure the model performance during the validation.

For the unsupervised approach, we are going to use clustering methods to see if any two businesses with different category tags can belong to the same cluster. Also, we will take random samples from different business – both from the same business tags and different tags – and observe the 95% confidence interval of topic similarity. Finally, we will use statistical methods to see if our embeddings can either accept or reject the null hypothesis: all restaurants with the same tags share the same topic distributions. To achieve this, we will construct the distribution of topic embedding for all restaurants with chosen tag, and use either Chi-Square or Kolmogorov-Smirnov(KS) test to see if our test statistics will show low p-values, in which case, we will reject the null hypothesis. If we reject this null hypothesis, we will use the confidence intervals to discover the optimal tags assignment for the particular business so that it minimizes the cosine distance within the newly assigned category.

3.2 Research Timeline

- October 22nd : First progress report
  We have already constructed LDA model and worked on the confidence intervals between randomly selected restaurants. This is how we discovered the topic model is not enough to generalize the reviews. By the second progress report, we will explore more generalized embedding such as stacked denoising autoencoder with generalization.

- November 5th : Second progress report
  During this week, we will work on the embedding using Word2Vec or stacked denoising autoencoder to finalize our embeddings. Also we will conduct supervised learning and generates precision/recall, ROC curve for various models. All the writing that involves supervised writing should be done by this.

- November 19th : Third progress report
  We will continue the research with statistical hypothesis test, namely KS test or chi-square. Essentially, we will discover 2 businesses that share the same business tags but result low p-value to reject the null hypothesis. It would be ideal if we could find the restaurants with different tags that show smaller cosine similarity of topic embeddings. Also we will finish generating the tables, graphs and results for the final paper.

- December 1st : Deadline
  We will focus on writing in the last 2 weeks.

3.3 Resources

- Pandas, gensim, nltk, numpy, scikitlearn libraries

- Business and review json files from the Yelp dataset

- Tensorflow and our generalized embedding systems with Stacked Denoising Autoencoder
- CDS server for possible category-merging to increase the data-richness of certain categories (might not be done for this project)

References

[1] P.Turney, *Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews.*

[2] X.Glorot, A.Bordes, Y.Bengio, *Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach.*

[3] P.Zhang, M.Komachi, *Japanese Sentiment Classification with Stacked Denoising Auto-Encoder using Distributed Word Representation.*

[4] H.Sagha, N.Cummns, B.Schuller, *Stacked denoising autoencoders for sentiment analysis.*

[5] D.Mimno, H.Wallach, E.Talley, et al, *Optimizing semantic coherence in topic models.*

[6] H.Wang, N.Wang, D.Yeung, *Collaborative Deep Learning for Recommender Systems.*

[7] H.Wang, N.Wang, D.Yeung, *Collaborative Topic Modeling for Recommending Scientific Articles.*

[8] F.Tian, B.Gao, D.He, et al, *Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves.*

[9] J.Huang, S.Rogers, E.Joo, *Improving Restaurants by Extracting Subtopics from Yelp Reviews.*