NLP Project Proposal

Varun Aggarwal, Vinayak Bakshi, Conrad De Peuter, Abhay Pawar

Columbia University, New York, NY, USA

Abstract

Our work intends to give insights to business owners (hotels first) by looking at what the customers are saying about specific parts of their experience on review sites (such as TripAdvisor or Booking.com). Time permitting we will apply our apply our approach to different verticals.

1. Aim

We broadly intend to do two things:

- Identify subtopics in the reviews and create a rating for these subtopics. We will identify what customers are saying about certain parts of their experience, i.e. service, ambience, ... on hotels, and come up with a rating for each of these from the text of a review. TripAdvisor has given subtopic ratings for categories so we will use these to verify our results.
- Predicting overall rating of a reviewer based on his review using the subtopic ratings from problem 1 as one of the features.

2. Motivation

The first subproblem is of immense use from the business owner's perspective as it gives them specific cues on how to improve their business. Identifying subtopics is an interesting problem especially when each vertical's subtopics are different. Eg. Hotels-(Value, Service, Location, Cleanliness ...). The methodology should be applicable across different businesses. Also, to infer the rating of that subtopic would require testing various features and methods. The second subproblem would be useful in inferring the rating of the business where there is no explicit rating given

3. Methodology

First subproblem: Various methodologies like Latent Dirichlet Allocation (Hoffman et al) or Latent Semantic Indexing have been proposed for mining subtopics. We would analyse how all these would work for our application and if we can improve upon these or come up with our own methodology. To predict the ratings for each subtopic we will first identify which part of reviews are talking about which subtopics, and then do sentiment analysis or feature engineering on these phrases. To verify our ratings on the subtopics we will use TripAdvisor's given subtopic rating.

Second subproblem: We will build upon the work of the first subproblem where we have identified subtopic wise rating and use them as features to predict an overall rating for a business. Time permitting we will try to apply our approach into different verticals on TripAdvisor.

For data, the Wang paper referenced below [4] provides a data set from TripAdvisor. If we do not find their data set sufficient we will scrape our own

4. Challenges

- TripAdvisor API not public and we would like different data we will need to scrape data from the pages.
- Reviewers don't discuss all subtopics in their reviews. How do we deal with reviews which don't discuss all aspects of an experience. The same problem exists for the specific annotated subtopics on TripAdvisor.com

5. Related Work

The following is related research in the field. Most relevant are the Honging Wang papers ([4],[5]) as well as [6]. We plan on building on this work and applying different methods to make our research novel.

[1] C. Papadimitriou, P. Raghavan, et all. "Latent Semantic Indexing: A Probabilistic Analysis." Journal of Computer and System Sciences. October 2000.

[2] D. Cai, Q. Mei, et al. "Modeling Hidden Topics on Document Manifold." Department of Computer Science, University of Illinois. CIKM 2008.

[3] M. Hoffman and D. Blei. "Online Learning for Latent Dirichlet Allocation." Neural Information Processing Systems, 2010.

[4] Hongning Wang, Yue Lu, Chengxiang Zhai. ?Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach" 2010

[5] Hongning Wang, Yue Lu, ChengXiang Zhai. "Latent Aspect Rating Analysis without Aspect Keyword Supervision". 2011

[6] Hao Wang, Martin Ester ?A Sentiment-aligned Topic Model for Product Aspect Rating Prediction". 2015