A Recommender System for Hotels Based on Review Factors

Zayd Al-Beitawi, Computer Information Systems

Mentor: Dr. Xuesong (Sonya) Zhang

Kellogg Honors College Capstone Project

Introduction

Every year, hotels receive thousands of detailed reviews from consumers. Consumers breakdown their experiences at the hotel by elaborating on a wide variety of factors including the quality of the check-in service, cleanliness of a room, availability of facilities, a welcoming atmosphere. Reviews impact consumer decisions as well as the areas hotel management chooses to improve upon to create a better experience for future customers. With large number of reviews constantly being posted online, it is difficult to pinpoint specific factors that customers or managers may be looking for regarding booking or fixing the hotel. We extracted 321 Yelp hotel reviews, parsed them into 2866 sentences, and categorize them using six hotel review factors adopted from previous research studies - service, price, location, cleanliness, room quality, and facilities. This trained model can be used to create a content-filtering recommender system to support the decision-making process of hotel consumers and management.

Hotel Reviews

Extremely vital to hotel management and development is understanding what exactly a future guest is looking for or is in need of during his or her stay. This allows hotel managers to implement crucial changes to the services and programs of the hotel, eventually "leading to higher customer retention" (Dolnicar, Otter, 2003). In their paper "Which Hotel Attributes Matter?", Dolnicar and Otter narrowed down the most important hotel review attributes by extracting over 173 attributes from over 20 hospitality, tourism, and business studies. They then grouped similar factors together and ranked them based on how frequent they were mentioned in all the studies. Upon evaluating the term frequencies in our Yelp hotel reviews dataset, we further narrowed down and combined the 13 attributes into 7 final attributes: Service, Price, Location, Cleanliness, Room, Facilities, and Atmosphere (Table 1).

Attribute	General Concepts
Service	Professionalism, service speed, friendliness, room service
Price	Value for money, deals, discounts, scams, frauds, coupons, overcharging
Location	Vicinity from parking, vicinity from Vegas strip (very prevalent topic in hotel reviews), vicinity from restaurants
Cleanliness	Stains, smells, cleaning services, insects, cleanliness of covers/bedsheets, cleanliness of bathrooms, messiness
Room	Size, noise levels, amenities (TV, PPV, Mini fridge), smoking/non-smoking, comfortability, views, bathroom amenities
Facilities	Fitness, pools/jacuzzis, spas, business centers, casinos, restaurants, shows/theaters, arcades
Atmosphere	Aesthetics, reputation, hotel themes/concepts, reputation of fellow hotel guests, decor

Data Preparation & Coding

We extracted 321 Hotel Reviews from the Round 12 Yelp Data Challenge Dataset, then parsed them into a total of 2866 review sentences using Python. Two researchers then manually coded each review sentence to 7 review factors (table 1). Once this was completed, the intercoder reliability was tested to ensure consistency and accuracy of the coding.

Model Construction

As a result of the efficiency and increased accuracy associated with Naive Bayes, we felt that it was the best potential algorithm to use for this text mining task. Once the training data was ready, we created a model using Naive Bayes to predict whether each review sentence is reflective of the concepts entailed by that review factor. To optimize the model, we applied text mining techniques including tokenization, filtering out stopwords, sample-balancing, stemming, N-grams, adding a dictionary for synonyms, and feature selection.

Results and Discussion

Through multiple rounds of testing, our model was able to achieve successful, predictive results. The best attribute was "Service", with an accuracy of over 99%. This is mostly like due to the fact that it was the most prevalent and elaborate review factor from our raw data set with over 700 relevant cases. It can also be concluded that the most talked about factor when it comes to customer hotel reviews are experiences related to the type of customer service received as well as the overall attitude of the staff and management team. Coming in a close second is the room attribute, to which our model was able to correctly predict over 98% of the cases during testing. Alongside the type of service a customer receives, the room quality and amenities seem to be a hugely discussed topic of interest when a review is written. Lastly, with an accuracy of over 97%, our model correctly predicted a majority of the facilities cases. Overall, our recommender model is extremely efficient and beneficial due to the fact that the most prioritized and prevalent review attributes from hotel reviews also have the highest accuracy rates when being tested by our model.

Variable	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha (Nominal)	N Agreements	N Disagreements	N Cases
Service	94.4%	0.856	0.856	0.856	2706	160	2866
Price	97%	0.783	0.783	0.783	2779	87	2866
Location	99%	0.866	0.866	0.867	2838	28	2866
Cleanliness	96.2%	0.811	0.811	0.811	2758	108	2866
Room	95.3%	0.799	0.799	0.8	2730	136	2866
Facilities	96.2%	0.78	0.78	0.78	2756	110	2866
Atmosphere	94.9%	0.472	0.472	0.472	2721	145	2866

The Proposed Recommender System

The purpose of this study is to develop a recommender system for a wide variety of hotel reviews based on the review factors finalized above. A predicted weight for each review factor will be established as the recommender system interprets each review at the sentence level. The significant impacts of this include automating the review classification process based on what the consumer is specifically searching for, as well as assisting hotel managers and employees with key business decision making that can be used to improve areas of question or concern.