BA 820 Team Hotel Reviews
Xiang Zhou, Chiebuka Onwuzurike, Maro Derhovanessians

Hotel Reviews Final Deliverable

Define the Problem

The topic we're diving into and researching was about the online reviews left by visitors from different hotels located in the US. Our objective is to perform review text analysis in order to provide key areas of importance and concern that affect review and sentiment score. The methods we used were NLP oriented, Rating distribution, Sentiment analysis based on reviews (both token analysis and sentiment methods comparison), and Clustering analysis.

Dataset

The dataset contains a list of 1,000 hotels and 10,000 reviews from 2014 - 2018 provided by Datafiniti's Business Database. The dataset includes hotel location, name, rating, review data, title, username, and more. Thus, we can use the dataset to compare hotel reviews by state dimension, and testing on sentiment scoring and natural language processing techniques. We mostly focused on reviews text/ title, reviews rating, reviews date, users city / province, hotels name and city.

Data Processing

Feature Engineering

The column names of dataset are kind of verbose and complex for implementation. And some columns are not required for exploration on the review text. Therefore, to design the data to be more efficient during use, we removed idle columns and renamed some to briefer words for easier readability.
Additionally, to help us better assess the relationship between words count and rating. We counted the number of words in each review.

Data Cleaning

Having cleaned data will ultimately increase overall productivity and provide a reliable dataset for later analysis. In the hotel reviews data we have, every row is unique and with no repetition. Additionally, there were only 3 observations with null values, and the null values were in 'title' and 'user.province' which were significant variables. So we decided to fillna() with "None" rather than drop the rows.

Preliminary EDA

Geographic Dimension

To have a basic understanding of the dataset, we took a look into the source of online reviews. Most reviews come from cities like San Diego, San Francisco and Seattle. Some from New Orleans, Atlanta, and Orlando. It provides us with an idea that people are more likely to book hotels from the west coast. It's reasonable to assume that west coast hotels have more attraction to visitors or travelers than middle and east coast cities.

Time Dimension

Based on the Month Dimension, most of the reviews occurred during the summer months, especially in July, August, and October. Based on the Year Dimension, most of the reviews from the time period between 2014 - 2018, with 2016 being the largest in volume.

Rating Dimension

According to the analysis, most of the review ratings for hotels happen between 4 and 5 stars, which means most reviews from hotels are actually high and positive.

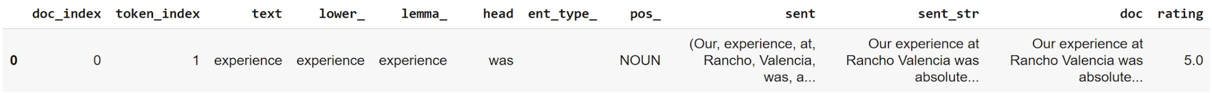
Token Analysis

SpaCy English Medium

We used spaCy to parse and tokenize our review data. Specifically, we made use of spaCy's Medium size English Core Web Modeling Pipeline ("en_core_web_md") because it would be well suited for written-text sourced from the web which our reviews could fall under.

Token Attributes

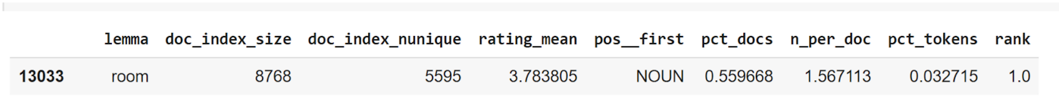
The attributes take from spaCy that we would be using would be 'doc_index', 'token_index', 'text', 'lower_', 'lemma_', 'head', 'ent_type_', 'pos_', 'sent','sent_str','doc'. The most useful of the list would be the doc and token index, lemma, entity type, and part of speech. Ultimately we would feed these attributes to a dataframe to calculator token stats.



	doc_index	token_index	text	lower_	lemma_	head	ent_type_	pos_	sent	sent_str	doc	rating
0	0	1	experience	experience	experience	was		NOUN	(Our, experience, at, Rancho, Valencia, was, a...	Our experience at Rancho Valencia was absolute...	Our experience at Rancho Valencia was absolute...	5.0

Token Stats

The dataframe 'token_stats' aggregates tokens together and calculates the meaning rating for the reviews they appear in, the amount of times they occur in a document, and the amount they occur in a unique document. This dataframe would be very instrumental in analyzing the different parts of speech of tokens, its frequency in the corpus and documents, and its association with different rating reviews.



	lemma	doc_index_size	doc_index_nunique	rating_mean	pos__first	pct_docs	n_per_doc	pct_tokens	rank
13033	room	8768	5595	3.783805	NOUN	0.559668	1.567113	0.032715	1.0

Token Filtering

We decided to look into what were the tokens for positive and negative reviews in a gradual level based on parts of speech. Thus, we look at the verbs, adjectives and nouns. To determine a positive token the mean review rating had to be greater than 4. For a negative token the mean review rating had to be less than 3.

Verbs

For positive token verbs, we found positive words such as love and enjoy. For negative token verbs, we found negative words such as smell, break, suppose. However an interesting similarity we found for both is they had words indicating they would share their experience, such as recommend, tell, say. Customers with a positive or negative experience were likely to share their opinion with others

Adjective

For positive token adjectives, we found expected positive words such as good, clean, great, nice, friendly. Likewise for negative token adjectives, we found expected negative words such as dirty, rude, poor, terrible, disappointed.

Nouns

Nouns would be a very important dimension to look at seeing as nouns do the action verbs and have adjectives describe them. Additionally they can be areas of focus that we can recommend to hotels. For positive token nouns, we found areas of focus such as staff, breakfast, pool, and restaurants. Alternatively with negative token nouns, we found areas of focus such as the carpet, toilet, smoking, stains, and management.

Sentiment Analysis

For sentiment analysis we looked at 3 different methods: Affin, TextBlob, Vader. Each has their own strengths and weaknesses. After getting sentiment scores for each method we wanted to compare how accurate each method was by defining their score by positive, negative, or neutral and comparing it to the ratings.

Affin

Affin typically gives a polarity score for each word between 5 and -5. We applied it to each document in the corpus. For Affin document scores greater than 5 we assigned it positive sentiment, less than -5 we assigned it negative, and in between neutral.

TextBlob and Vader

TextBlob and Vader typically give a polarity score for each word between 1 and -1. We applied it to each document in the corpus. For TextBlob and Vader document scores greater than .2 we assigned it positive sentiment, less than -.2 we assigned it negative, and in between neutral. We decided between .2 and -.2 because it aligns with the one-fifth neutral sentiment assigned to ratings of 3.

Comparison / Accuracy

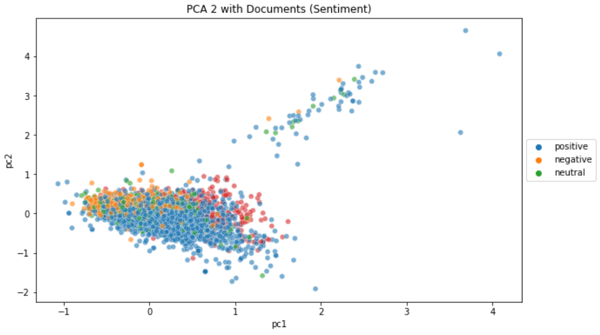
We set the sentiment rule by classifying ratings > 3 as positive reviews, ratings = 3 as neutral, and ratings < 3 as negative reviews. To measure the accuracy of each sentiment method, we checked if their results match the sentiment rule. And below is the comparison between these sentiment methods, which comes out Vader received the highest accuracy. It assists us to weigh more sentiment analysis by Vader.

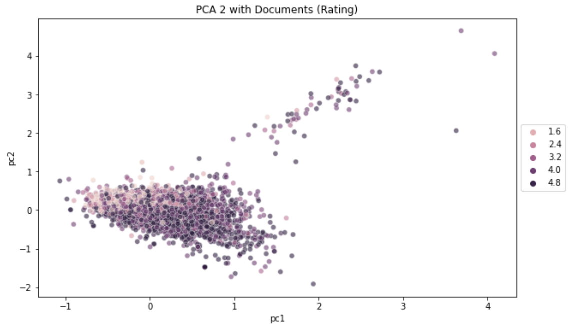
Sentiment Method	Afinn	TextBlob	Vader
Mean	16.09	0.2893	0.7951
Accuracy	0.7025	0.7207	0.7956

Dimension Reduction Analysis

PCA2 with Documents

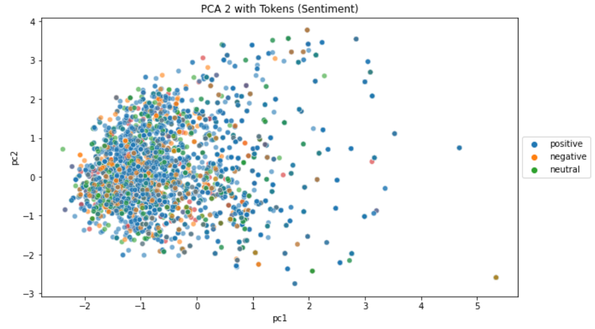
PCA2 with documents vectors seems to have a more condensed 2 dimensional representation. There seems to be a clearing area of rating and sentiment clustering when using the documents.

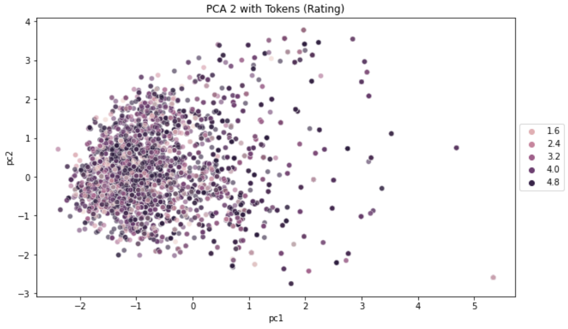

PCA 2 with Documents (Sentiment)


PCA 2 with Documents (Rating)

PCA2 with Tokens

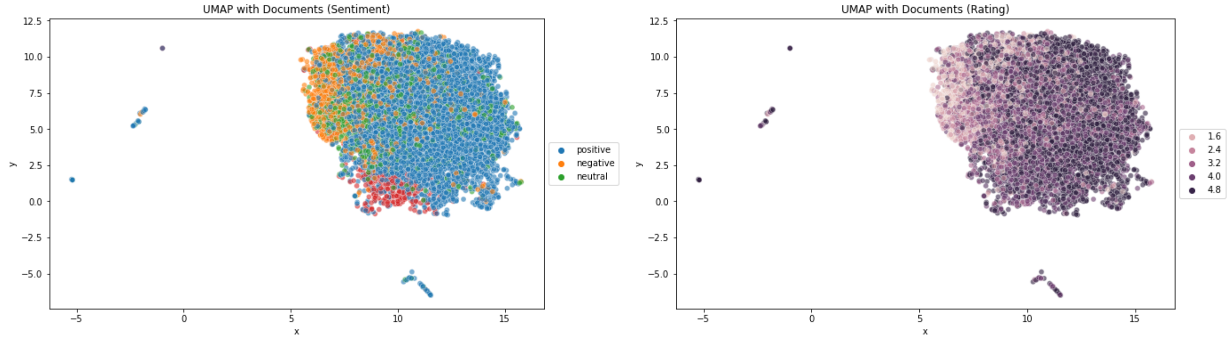
PCA2 with token vectors seems to have a more spread out 2 dimensional representation. No clear clustering based on rating or sentiments.


PCA 2 with Tokens (Sentiment)


PCA 2 with Tokens (Rating)

UMAP with Documents

UMAP with document vectors seems to be condensed as well just like PCA2 with very few outliers. However there seems to be more clustering of the sentiments and and rating. This could make sense because UMAP is able to explain more of the variance than PCA2.



Conclusion

We started off with a simple objective to perform review text analysis in order to provide key areas of importance and concern that affect review and sentiment score. Through preliminary data analysis we found our reviews were skewed towards warmed locations such as California and Florida, skewed towards higher ratings (4s and 5s), and written during the warmer months between 2014-2018. Reduced the dimensions and visualized our review text in a 2 graphical representation using UMAP and PCA2. Additionally we found our sentiment analysis to be very accurate in terms of classifying review text based on positive, negative and neutral sentiment. Lastly and most importantly we provided token analysis after aggregating and ranking our tokens. Ultimately found key areas of focus in the token nouns. Positive reviews contained nouns such as staff, breakfast, pool, and restaurants. Hotels need to focus on these areas as they are favorably spoken about the most. Negative reviews contained nouns such as carpet, toilet, smoking, stains, and management. Hotels need to be cautious of these areas in order to not get bad reviews and stand out above the crowd. Regardless, people who wrote reviews are very likely to recommend, tell, or say something to others which could drive or decrease future business, thus supporting the importance of review text analysis.