Sentiment Analysis of Hotel Reviews

POLITÉCNICA

Intelligent Systems: Natural Language Processing

Mónica Maldonado Olivares

1 PROBLEM TO SOLVE

The chosen dataset collects 20000 reviews written by travellers about the hotels that they have been stayed. This dataset also contains the evaluation that they the customers gave to each hotel between 1 and 5.

For this analysis the objective is to analyse the reviews in order to discover which are the most repeated words and also try to extract information about its structure. The results could be useful if someone with a hotel reservations webpage wants to add or improve filters for the user who are looking for something specific. Also, I can use these results to discover the most repeated topics. The other objective with this project is to build a model capable of predict whether a review is going to be positive or negative.

First, I have applied some changes in our data to be able to work with them. I have removed those reviews which have been written in a different language instead of English and I have processed the data normalizing and cleaning it.

- Convert it to lowercase
- Remove numbers
- Remove punctuation
- Remove Stopwords: this is to remove all the words that doesn't provide meaning like "its" or "he".

2 EXPERIMENTS DONE

For this analysis the following packages were used:

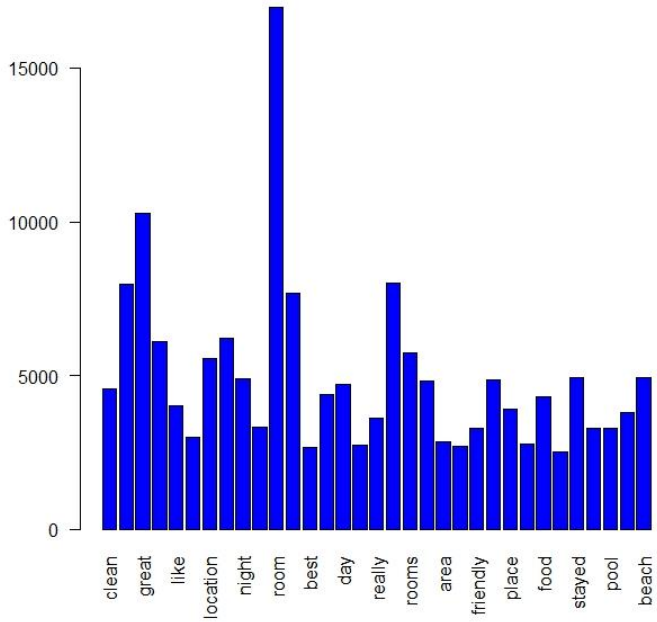
- "cld3": for detecting languages
- "tm": for text mining
- "wordcloud"
- "quanteda": for managing and analysing
- "quanteda.textmodels" and "caret": for the predictions

2.1 Word analysis

In this analysis the goal is to represent the frequency of the most repeated words using a bar chart. After remove some elements like numbers or punctuation and clean the reviews to make them processable, the function "TermDocumentMatrix" has been used to obtain a matrix where in the rows it has all the words that are in the reviews and each column represent each review. However, it was necessary to reduce the size of the dataset because it wasn't possible to build a matrix as big as that one. That's why I have done a random sample of the reviews which size was 8000, i.e., I've selected 8000 reviews.

	Docs							
Terms	1	2	3	4	5	6	7	8
audio	1	0	0	0	0	0	0	0
august	2	0	0	0	0	0	0	0
available	1	1	0	1	0	0	0	0
away	3	0	0	0	0	0	0	0
awful	1	0	0	0	0	0	0	0
bad	1	1	0	0	0	0	0	0
balcony	1	0	0	0	0	0	0	0
bar	1	1	0	0	0	0	1	0
bars	2	0	0	0	0	0	0	0
bathroom	1	0	0	0	0	0	0	0
bay	1	0	0	0	0	0	0	0

Once I had the matrix. I selected the words with a higher frequency, in this case more than 2500 instances to represent only the most repeated words. There were some words between the most repeated ones that didn't give us information so I removed them and the following graph was obtained.



The words that were also very used but I have considered as not useful are: "goes","got","hotel","hotels","amsterdam","looked","looks","dont".

The wordcloud is also been done and with which it is possible to obtain a global view of the frequency of the words.

2.2 Sentiment classification

First, I have changed the values of the rating to have only two values: “bad” or “good”, 0 and 1 respectively and I have transformed the clean reviews into a corpus object. A corpus object is a collection of text documents, in this case, a collection of reviews.

Then I have added it a docvar with the rating of each review and I have split out the reviews into two sets: training set and test set, the first one is the 80% of the data and the other the 20%.

Hence, in order to avoid having a lot of reviews with the same rating if they were sorting, I have done a random sampling of the reviews and, after tokenized the reviews and create a document feature matrix of it, I have created the training set. The test set is made up of the remaining reviews.

There were to alternatives to build the model: using the Naïve Bayes algorithm or SVM algorithm. I have used both.

Naïve Bayes: Multinomial

	Reference		
Prediction	0	1	Accuracy=0.92
0	547	139	Sensitivity=0.68
1	252	4186	Specificity=0.96

Naïve Bayes: Bernoulli

	Reference		
Prediction	0	1	Accuracy=0.85
0	306	266	Sensitivity=0.38
1	493	4059	Specificity=0.9385

SVM: Uniform

	Reference		
Prediction	0	1	Accuracy=0.90
0	493	190	Sensitivity=0.61
1	306	4135	Specificity=0.95

SVM: docFreq

	Reference		
Prediction	0	1	Accuracy=0.90
0	472	168	Sensitivity=0.59
1	327	4157	Specificity=0.96

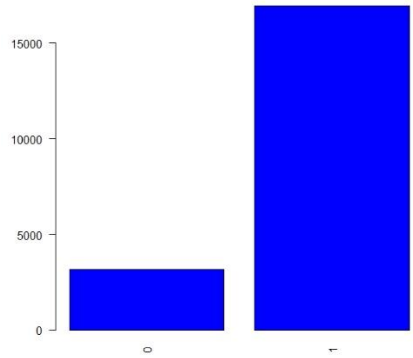
3 ANALYSIS OF RESULTS

In the section 2.1 I have done a simple analysis of the reviews with the aim of understand more the topic of our data besides to know which are the more frequent words that the customers use. As can be seen, the most repeated words in the reviews are the location, the hotel’s

facilities and some important services like the cleaning, the food and the personal. The output of this analysis is something totally expected and it doesn't give us more information.

Regarding the models that I have build in the section 2.2 the best model would be Naïve Bayes: Multinomial. However, as can be seen in the confusion matrix, the four models predict pretty well the positive reviews however, as far as the negative reviews the result are not accurate.

In the following graph it can be seen that the difference between the positive and negatives reviews are very significant and that's why I have arrived to the conclusion that that difference has impeded the algorithm from learning correctly from the data.



Therefore, I think that with a data with a less difference between the number of reviews with different rating I could obtain a better model.

4 GITHUB

In the following link are the code and data used for this analysis:

https://github.com/moni3889/NLP_Assigment