UNIVERSIDAD
POLITÉCNICA
DE MADRID

Federico
Rodigari
2020/21

Natural Language Processing

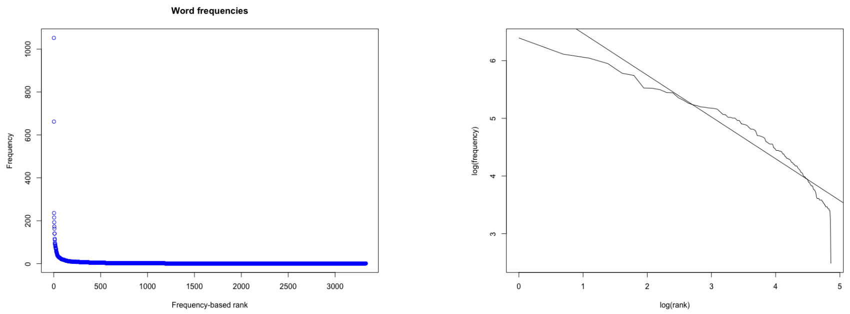
Sentiment Analysis of Tripadvisor Reviews

Abstract

With an increasing number of reviews over last years, the field of natural language processing plays an important role in helping companies setting the right marketing strategy and discover, in a scientific way, its strengths and weaknesses. Reading all the reviews and to understand what are the opinions of customers can be a complex operation if you are faced with a large number of reviews. Here is where *sentiment analysis* comes in. In this project data are scraped from TripAdvisor tanks to the work done from Giuseppe Gambino on Github and different analyses of text documents at a word level are performed. The project terminates with a sentiment analysis made possible from two different r libraries: syuzhet and sentimentr. The scope of the work is to support Hotel Lac Salin owners in their marketing strategy.

Preliminary operation

In the firs step of the process the scraped file containing the reviews is imported as a dataframe in R and transformed in corpus in order to perform the operations. After that, a term document matrix is produced in order to describe the frequency of terms that occurs in the collection of documents. At this point the Zipf's law is tested in order to ensure the reviews' consistency.



(a) Word frequencies (b) Zipf's law

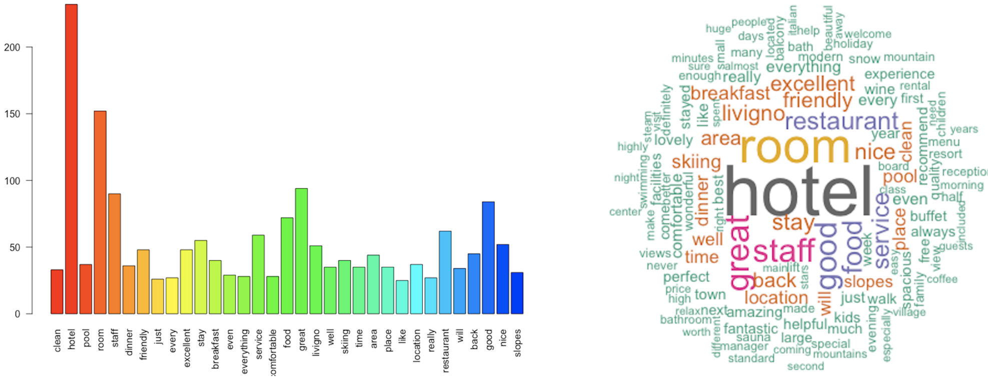
Thanks to this, we can claim that most frequent term can be ignored since they are probably poor discriminators for the scope. Consequently, stopwords for the English dictionary provided from the library **"tm"** and some othe useful words such as *"also"*,*"lac"* and *"salin"* were dropped. Another refinement is done in order to get a more precise analysis, in particular the words *"restaurants"* and *"rooms"* are replaced with their singular and *"witresses"* is replaced with *"weiters"* to more easily identify opinions about the hotel staff.

Analysis

Once the final dataframe was built, a first analysis was performed exploring monograms and bigrams.

Monograms

Monograms analysis account for single words, for this reason first relevant plot to display is the barplot of the single words that are present in the reviews. The graph (1a) allows to see the most relevant cited worlds giving an overview of the most discussed topics in the reviews. A better representation of this situation is given from Word Cloud (1b).



(a) Word count

(b) Monogram' WordCloud

Bigrams

A second analysis is performed to understand the relationship between words. For this purpose we exam bigrams that are pairs of two consecutive words. Those operation are supported by the library **"RWeka"** and allowed us to plot an interesting Word Cloud displaying the most frequent word associations. In addition to this, a barplot displaying the bigrams' frequencies sorted is computed in the code. Trigrams were also explored but not reported since the do not provide any further useful information.

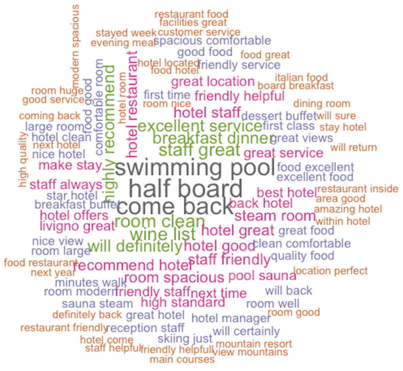


Figure 1: Bigrams' WordCloud

Sentiment Analysis

At this point, we have a clear understanding of the reviews and we want to determine the customer emotional reaction towards the hotel. This attitude is provided as a judgment in the review and can be a precious information or feedback for the hotel owners. Thanks to **”syhouze”** we calculate the presence of eight different emotions and their corresponding valence in a text file. Those emotions are *”anger”*, *”anticipation”*, *”disgust”*, *”fear”*, *”joy”*, *”sadness”*, *”surprise”*, *”trust”* and the occurrences of them result in a positive or negative score which difference gives the overall score.

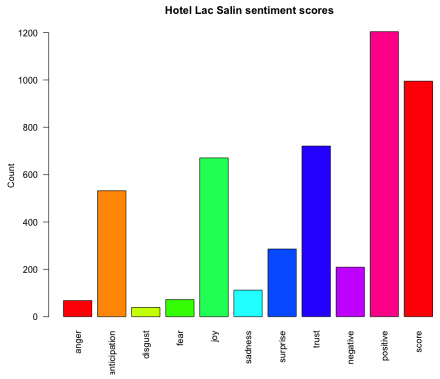


Figure 2: Sentiment scores

Although this analysis can be significant, it does not take into account the presence of negation preceding the word, that can drive to erroneous conclusions. This problem was solved trough the library **”sentimentr”**. The library provides the function *”sentiment_by”* that approximate the sentiment (polarity) of text by grouping variables giving a more reliable result.

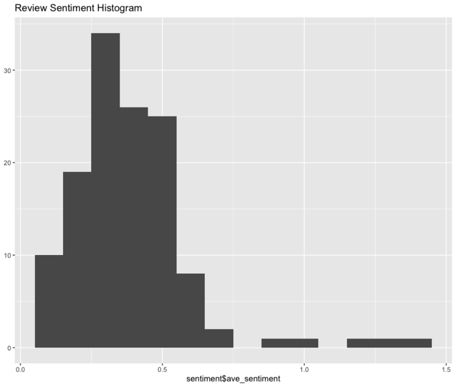


Figure 3: Review Sentiment Histogram

Evaluation

From the first analysis performed, it resulted as expected, that the most cited words in the reviews are *"hotel"* and *"room"*. This do not provide any relevant information and could have been added to the stopwords. However, different positive words occurs in the reviews such as *"great"*, *"good"* and *"comfortable"*. Other relevant discussed topics are *"staff"* and *"food"*. From the bigrams' Word Cloud can be seen that customer are often engaged in writing about the hotel "half board", *"swimming pool"* and *"wine list"*. Those could be the strengths of the hotel and the management have to take them into account. Along with this, some positive opinion are displayed: *"come back"*, *"staff great"* and *"room clean"*. From the sentiment analysis, how can be seen from graph (2) the overall score is very high meaning that most of the reviews are positive and they show elements of trust, joy and anticipation. The more reliable analysis made by **"sentimentr"** clarify the result of the previous identifying all the reviews as positive. This probably leads to the fact that **"syhouze"** has not considered some negations before words that were classified as negative.

Summary

Overall, the work done shows that the English reviews gathered from TripAdvisor in the period August 2007 - February 2020 of "Hotel Lac Salin" have a positive score, meaning that the hotel is very well regarded by its customers. The future work will be to take into consideration different languages reviews and to expand collection of data also to important players such as Booking.com. Another interesting work could be addressed by considering the reviews of an entire touristic village to support the development of his tourism based on the reviews of its visitors.