Harshit Bansal & Herbert Luiz Botelho Dias

ID 300303010 and ID 300286075

Abstract

This report contains the analysis of a explanatory study which goal is examining the quality of the medical facilities in Canada based on the customers reviews.

Project REport

Special Topics in Data analytics  
Instructor: Samuel Otim

**Part 1: Discovery (RateMd website)**

1. **Learning the Business Domain**

The medical environment is upside down after the advent of COVID 19. In these challenging times, many new concerns showed up and were enlightened by these news variables; we bring a slightly different analysis. While most analysts focus on the doctors, we approach the facilities to understand how people evaluate these places. Since the demand for these places is affected by the current scenario, the requirements for people attending these locals have risen.

1. **Framing the problem**

This project's initial purpose is to examine Canada's quality of medical facilities based on their reviews. However, we turned our analysis to discovery, for example, independent of which word's classification has more impact in each review based on its weight (coefficient). This task could be tricky since there is no structured data, and the content of facilities reviews can be hard to transform into a reliable data frame. Facing these problems, we study the data behaviour in the web site before the extraction and if we could find any visible pattern to give some light to overcome these initial challenges.

After some scraping attempts, we got visualized some patterns that could not affect our analysis. For instance, there is no classification among the facilities; usually, the name is an indicator of its specialty. And we decided to frame our goal and the problems that we could solve. The list below will show these findings

* What is the overall medical facilities quality in Canada?
* Which are the sentiment present in reviews?
* How can this sentiment (words) affect the review?
* Can reviews define the quality?
* Exploratory analysis is suitable for this data
* Which is the best model for the analysis ( We found Logistic Regression)

The findings above will guide the analysis to be more assertive regarding the facilities' quality. Using the text, can we provide suggestions for potential changes in the facilities we might need.

The outcome searched in this project are words that represent negative and positive sentiment. Also, the model test presents an analysis in a random review sample.

1. **Initial Approach**  
     
   Based on the observed scenario, our analysis approach will follow the steps below defined in the proposal.   
   This report will provide screens shoots of each step performed in Python.  
   * Scrap the data from RateMD
   * Create the DataFrame
   * Preprocess the Data

* Preliminary Analysis
* Formating Converting Text
  + Implement classification
  + Build Models ( Bag of Words)
  + Extract Features
  + Perform Sentimental analysis (Logistic Regression Analysis)
  + Perform tests
  + Display and analyze the results

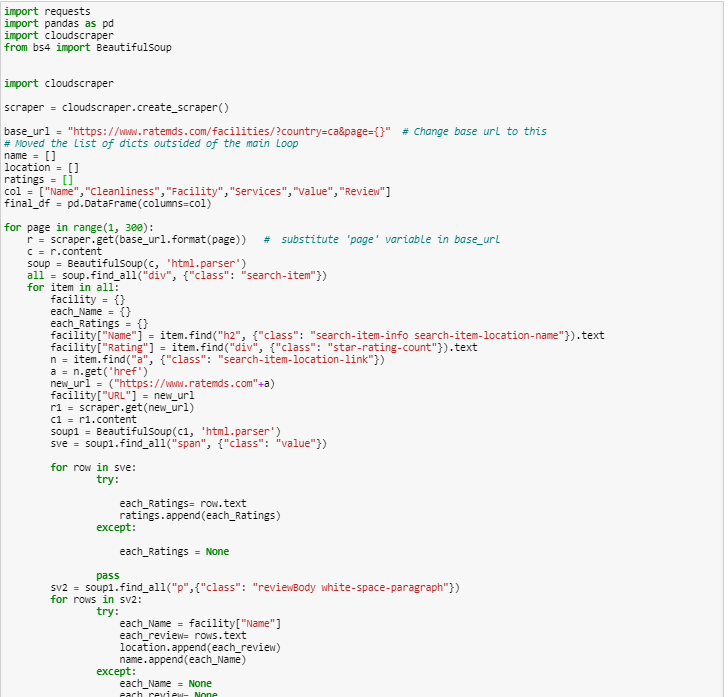
1. **About the Data Set**

To conduct this analysis, we scrapped the facilities reviews from the RateMd web site, <https://www.ratemds.com/facilities/>. This scrapping process covers the facilities in Canada without a region filter. We opted to use all data because we observed a trend to repetition in the ratings, and in this case, the size of our dataset needs to have at least four thousand reviews. We gathered around seven thousand facility reviews in all of Canada and defined the headers described below.

First Version of the Scrapped Data.

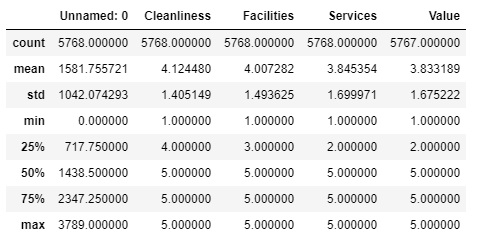


Follow below the script created to scrap the data from https://www.ratemds.com/facilities.  
We are using a cloud scraper to bypass the Cloudflare antibot agent. We have a base URL that is iterated each time we must access the next page. We get into a page, fetch the facility name and total reviews for that facility. While scrapping, we encountered significantly less recent data; hence, we could not perform our analysis over the countermeasure's facility might be taking or should take in this pandemic. But we went further and scrapped all four ratings in a list, then divided that list into four ratings.

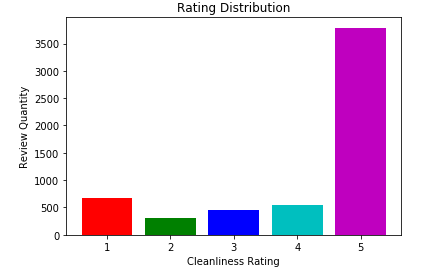
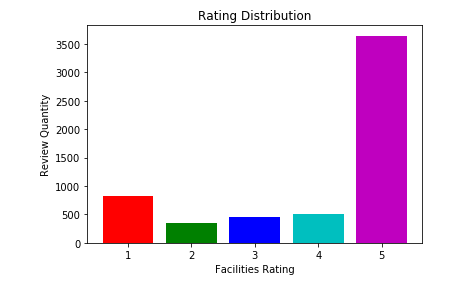
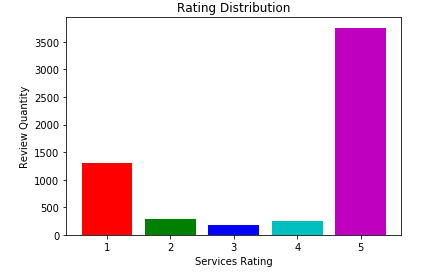
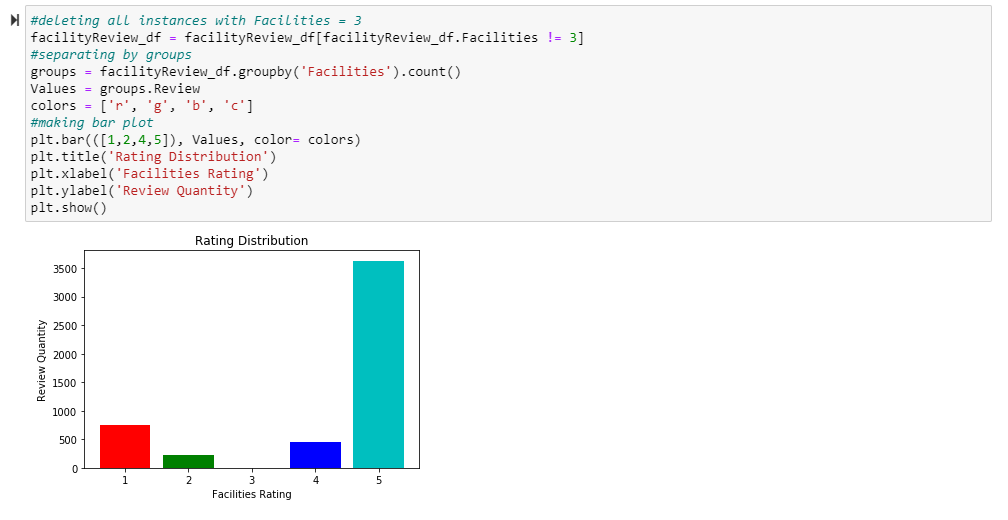


1. **Challenges**
   * Identify the proper approach to the sentiment and bring them into classes; we extracted the features gathered with the analysis and used models such as BOW (bag of words) to remove them and build text classification models (TF-IDF).
   * As mentioned in the previous section, dealing with the data was challenging because it tends to disproportional for positive reviews. A considerable percentage of the reviews are positive, so we should find a method that could bring the model's reasonable accuracy and make this prediction meaningful.
   * The Scrapping of review details was complex to build.
   * Get the data in RateMd. As a security method, the RateMd web site has a block that avoids the first attempt of Scrapping. To bypass it, we used CloudScrapper(bypass Cloudflare's antibot page). Before that, we made several attempts without success.
   * Tuning the models to make the prediction meaningful.
2. **Analytic Techniques Used in this project and Outcomes**
   * Exploratory Analysis
   * Scrapping
   * Text Analytics and NLP (Natural language processing)
   * Text Mining Classification
   * Perform Text Analysis Operations using NLTK (we started with spark, but we got problems and alter to NLTK)
   * Bag of Words
   * Logistic Regression
   * Lemmatization (reduces words to their base word)
   * Tokenization (word, sentences)
   * Frequency Distribution
   * Eliminating Stop Word
   * Sentimental Analysis

We performed an exploratory analysis with a text-mining classification model using a bag of words and logistic regression. We are using a different approach rather than the one that was mentioned in the proposal. As we opted for the logistic regression, we should separate the reviews in good or bad based on the rating score. Besides, it is interesting to know each word's coefficient, specifically for each group (Cleanliness, Services and Values). But first, we extract a big picture of the sample.

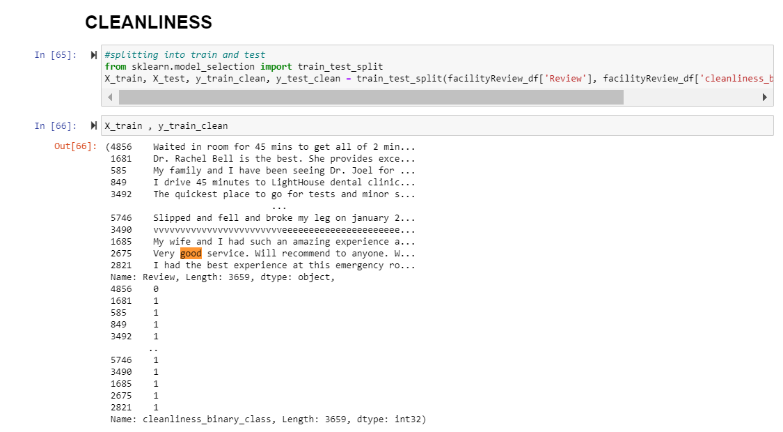


We can observe the mean for each category got a high mean. However, services and Values have a slightly lower result than others. In this case, the model can clarify the results above.  
The prints below can prove that we have a disproportional amount of 5 stars reviews.

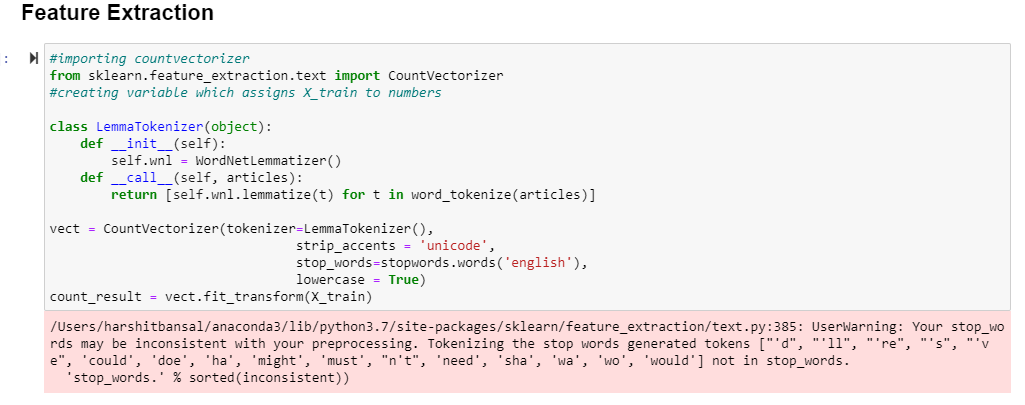


To perform logistic regression, we need to delete the ratings with the value of 3, due it is neutral on a scale from 1 to 5. For this analysis, the number 3 does not offer any sentimental insight and can be discarded to perform logistic regression. This step was repeated for each category.

The creation of a binary class is necessary to divide the groups into good or bad reviews, according to the prepared scale removing the number 3 as mentioned above.

  
Splitting the data into 75% train and 25% test data, we will work on cleanliness data to get the negative and positive words corresponding to the cleanliness ratings and then we will check that for the rest of the ratings. However, the ratings for services and value are not that different as we saw in the above graphs and from the describe () method.Using NLTK.

Lemmatization was used to reduce similar words and improve analysis accuracy. Terms such as analog and analogical will enter in the same group.



Measuring the model accuracy



The AUC for this model is particularly good, and it says we have 88% of correctly classified instances.

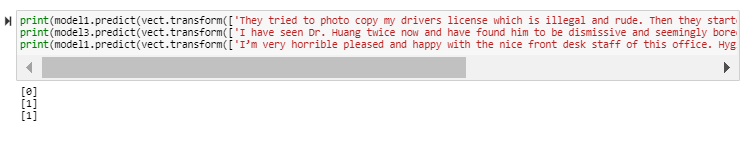
We will be able to pinpoint specific words with a high impact on rating sentiment. This information could be an asset for Facilities (if used for their entire internal dataset — not this not so big sample we have). For example, the word "rude" likely has a negative coefficient, pushing our classifier to label that review as bad. The term "great" likely has a positive coefficient, pushing our classifier to label that review as good. The goal is discovery. We do not know what interesting words may appear to impact sentiment, and that is the fun part. Does the word "great" outweigh "horrible" and net positive or negative? Many questions like this will be answered with our model. Follow below the coefficient for cleanliness, services, and value, respectively.



1. **Test results and Managerial Implications**

Since the whole data is gathered, divided, and processed, it is time to run the model and analyze if it presents the desired result. Fortunately, the model worked well and brought the expected outcome. Our model could indicate if each word has the most negative impact in each review section. For instance, the word "rude" is the champion of negative sentiment present in all categories, evencleanliness. This reinsures that Canadians are polite, and kind peoples and rudeness are not welcome. The opposite feeling also follows a pattern for services and facility, "friendly" is the word with the highest coefficient among the favourable terms. If we find these two words in the same sentence, the model will balance the coefficient, bringing our results' desired accuracy.

Follow below an example of how the model works.



First, we need to consider that zero represents a negative review and one a positive. As we can see in line three, we have an excellent example of how our model is working. The sentence starts with horrible, which is the top 3 negative words. However, the coefficient of the other good words that complement the review, such as please, happy, and nice, made the model provides the correct result and classify the review as positive.

We could observe that some words are indicated explicitly to a different review area as an exciting discovery. For instance, in value, the word "hour" a top 5, while it is inexistent in the other groups.

The results presented below can provide indispensable business insights to a determined facility. As implication rudeness, and inadequate facility quality needs to be closely observed. The staff training for those who have bad reviews needs to be intensified, and for the facilities which contain horrible in their reviews, I advise check-up in the whole structure, independent of the size. Of course, huge hospitals tend to have more reviews and apply these results to them would be more challenging than using for small facilities such as walk-in clinics.

1. **Tools and libraries you need**

For our project, we used these libraries and tools

Sklearn

Pandas

Numpy

Beautiful Soup

SeaBorn

CloudScrapper

Beautiful Soup

MatPlotlib

Count Vectorizer

NLTK Tokenize

NLTK WordNetLemmatizer

TfidfTransformer

roc\_auc\_score