Mathematics for Analytics
MA/ MSc in Analytics (Class of 2025)
Tata Institute of Social Sciences, Mumbai
Assessment 3 | Maximum marks - 20

Submitted By:
Vivek K S (M2023ANLT019)

Text Analysis of Google Reviews on Tata Institute of Social Sciences, Mumbai

In this project, my attempt is to analyze the Google reviews made by the public on the Tata Institute of Social Sciences, Mumbai campus. The rationale behind deciding to analyze the google reviews of TISS Mumbai is particularly to understand the public perception of TISS as an institution. As a student of TISS Mumbai, I was eager to know how the public considered this institute on academic grounds and other factors. Google map reviews is a public platform where anybody can make a comment without any restriction or entry barriers. Hence, I believe this would give the public the possibility to express their opinions freely, whether it's a negative or positive opinion. The opportunity to study these reviews would give us a clear picture on how people perceive TISS Mumbai on various factors, and what are the major remarks made.

The reviews were extracted from the google map location of the TISS campus by web scraping techniques using Python. An elaborate process was required to scrape out the google reviews (specifically the ones with text) into a csv file. However, we would not delve into the details of the web scraping process at this point as this is beyond the scope of this particular assignment. As on 13-10-2023, it was observed that there were a total of 787 google reviews on TISS Mumbai, and out of them, 304 had text content in their review, while others have given their rating without any comments. Hence, finally after the completion of web scraping, I could generate a csv file with 304 rows and two columns, namely, the reviewer name and review text.

The csv file generated above was imported to R for Text Analysis of reviews. The next step was to analyze the total length of text content in Google reviews to understand the average length of a review made. It was identified that the average length of a google review on TISS Mumbai is of 116 characters.

Then, I had to perform the pre-processing steps to clean the data and prepare for analysis. In this context, the review column was first converted to a volatile corpus and then passed on to the cleaning function. The cleaning function would perform the steps of transforming content to lowercase letters, removing stop words, removing punctuation, removing extra white spaces and numbers. The stop words list was customized earlier to include the words like “Tata”, “TISS”

and “Institute” as they would be frequently used in the review statements but would not add much significance to the analysis. The steps of preprocessing were performed in the exact logical order to achieve the desired results.

After the cleaning process, it was observed that there are 1108 distinct words used in the reviews. Now, the next step was to identify the frequency of these words used in reviews and to understand the words of highest frequency. The rowsum( ) function was used to calculate the frequency of words used. Finally a new dataframe was created with each word and its frequency. The top 10 occuring words were displayed as below:

S. No.	Word	Frequency
1	social	114
2	best	92
3	campus	83
4	place	66
5	great	49
6	india	48
7	sciences	47
8	good	42
9	one	42
10	education	39

After the creation of the word frequency table, the next step was to generate a word cloud based on this data. The package - wordcloud2 was used to generate the word cloud. The final output of the project was a beautiful word cloud generated out of the google reviews made by the public on TISS Mumbai Campus. It was interesting to note that the public considers TISS as one of the best campuses for social studies and a great place for education as these were the most repeated words in the reviews.

The word cloud generated was published to the RPubs and can be viewed using the link given below: https://rpubs.com/vivekks/textminingproject

****