NLP Project 8

Mitja Kärki (group leader),
Eetu Laukka, and Syed Muhammad Shahrukh
https://github.com/syedmuhammadshahrukh/nlp-sentiment-analysis

***Abstract* - All papers must include an abstract and a set of index terms. The Abstract and Index Terms text must be 10 point Times New Roman, fully justified and contained within one paragraph. Begin the Abstract with the word Abstract - in Times New Roman italic. The entire Abstract should be in bold. Do not indent. Use a standard dash after the word "Abstract." Do not cite references or use abbreviations in the abstract. It should be approximately 200 - 275 words.**

Index Terms - About **four**, placed in **alphabetical** order, key words or phrases that are separated by commas (e.g., Camera-ready, FIE format, Preparation of papers, Two-column format). Italic for the label "Index Terms"; otherwise, regular font.

INTRODUCTION

This project aims to investigate the sentiment and test various architectures for comprehending sentiment polarity in London hotel dataset**.**

The growth of online bookings has enabled multiple travel companies to rise and to allow hotels around the world to keep information readily available to everybody on the internet. These hotel reviews are naturally an important source of information for travel planners. Thus, this information is also crucial for hotel managers. The understanding of this information can take time if done by hand, thus sentimental analysis system of this information can greatly benefit hotels and their managers to improve their business and to market better [1][2].

This opinion-rich information is now very popular and there is a lot of it. It is not always used to its full potential and thus it provides big opportunities for different approaches of opinion mining and development of systems and frameworks for it. Many different statistical approaches have been developed, but the full understanding of natural language by computers has not been achieved yet, this calls for more research for it [2][4].

Zvarevashe and O. O. Olugbara provide a framework for opinion mining for hotel reviews for sentimental analysis [2]. Our approach uses very similar strategy as this framework but we go a step further with our novel approach to then continue with the sentiments and analyze the first and second order statistics of the user's rating. After which, we identify the hotels that have low standard deviation and the ones with high standard deviation.

This approach allows us to see whether the high variation of sentiment happens in the higher rated hotels or lower rated hotels. After this, we analyze these selected hotel reviews and look at the content of those reviews by using different techniques to find out the categories of the reviews. Both for negative and positive reviews. This gives us good overview of which topics arise most on negative and positive reviews. This approach would give a very detailed source of information for hotel managers to focus on improving the topics that arise in the negative reviews and to know what is working from the positive reviews. We use LDA (Latent Dirichlet Allocation), empathy analysis and we also create our own categorization by using Wu-Palmer Similarity to check word frequency within popular hotel review categories. This overall shows a comprehensive analysis of review and could be used by hotel owners and managers to gain significant insight into their business.

THE PROBLEMS

The team started by analyzing the dataset available from Kaggle [6]. This gave insight into what type of data we are dealing with and the structure of the data. The first task was to sanitize the reviews and then put it into SentiStrenght [7]. Then the output of that sentimental analysis would be fed into the program which will evaluate the correlation of the overall sentiment score of each review with Pearson correlation coefficient. Then the goal is to calculate the first and second order statistics of a user's rating, from which we would take only certain hotel's reviews for further analysis by choosing a threshold of high variance. These reviews then would be visualized with a word cloud. This part included our problems from 1 to 5. The latter problems consider the analysis of these chosen reviews.

We will create five topics from these reviews for both negative and positive subclass. This is one through LDA. These topics can then already be assessed with the word cloud to look for similarities. This is already comprehensive result if such findings are to be found. After this step, we create empathy categories for these reviews which we can then compare between the LDA topics. This further confirms our results and findings. Finally, we would create our own ontology construction to fit words for popular hotel review categories such as: price, service, parking, room, location and food. This was done by looking at the Wu-Palmer Similarity between the words and the categories to

figure out the frequency of each categories in both negative and positive subclass.

THE DATASET

The dataset was a manually labeled dataset available from Kaggle [6]. This dataset contains reviews of the top 10 most expensive and least expensive London based hotels with hotel locations, reviews and user's rating. The dataset was built by crawling a leading travel portal by PromptCloud. The dataset contains these values: Property name, Review rating, Review Title, Review Text, Location and date of the review. This dataset is free to use for public exhibitions such as this project, since the dataset is under the license: CC BY-SA 4.0. The data is no longer updated or even up to date but since this is a student project this does not matter since this is only for learning purposes. The last update of the data is from 2018.10.23. We have included a snippet of the first two rows of the dataset in figure 1.

Property N...	Review Ra...	Review Title	Review Text	Location O...	Date Of Re...
Apex London Wall Hotel	5	Ottima qualit♦ prezzo	Siamo stati a Londra per un week end ed abbiamo alloggiato in questo ottimo Hotel prenotato da amici...	Casale Monferrato, Italy	10/20/2012
Corinthia Hotel London	5	By far, my best hotel in the world	I had a pleasure of staying in this hotel for 7 nights recently. This hotel was perfect in every way...	Savannah, Georgia	3/23/2016

FIGURE I
Snippet of the dataset

METHODOLOGY

Main tools

For the majority of the sentiment analysis, Python 3 was used since it is a commonly used tool within natural language processing and provides a variety of related software libraries. These third party libraries can often be installed using PIP (Python package installer). To present the work in a neatly organized manner, a tool called jupyter-notebook was used [8]. It provides a web browser based code and text editor environment to neatly present the source code, its execution and results. Jupyter Notebook offers also slight GUI to show headers for example and showing the results. But it is very barebones. The workflow is briefly explained in the following subheadings and visualized in figure 2. The jupyter-notebook file is available at the project's github page: *https://github.com/syedmuhammadshahrukh/nlp-sentiment-analysis*. To review the code, jupyter-notebook needs to be installed. When running the code, the reviewer might need to install some Python packages as told by the UI (pip3 install <packagename>).

Data cleaning

The analysis began with cleaning the data in various ways for processing. The steps included converting the data to lower case, tokenizing it and removing punctuation and stopwords.

Sentiment scoring

To determine the positive and negative sentiment scores for every review, SentiStrength was used. After that, the data was imported back to Python where the overall sentiment was calculated.

Statistical analysis

Next tasks were to calculate a correlation coefficient and first and second order statistics grouped by hotels (mean and standard deviation).

The following task was to select a standard threshold for creating an ambiguous class. Furthermore, the hotel was also added to either negative or positive ambiguous subclass depending on the user ratings.

Topic distribution

Topic distribution of the positive and negative subclasses was determined utilizing LDA. Then, Empath client was applied to the LDA model's negative and positive subclasses [9]. The overlapping ratios between Empath categories as well as Empath categories and LDA was calculated.

Categorization

The words of the whole data were categorized into six standard categories. A histogram was again plotted of the frequency of the categories.

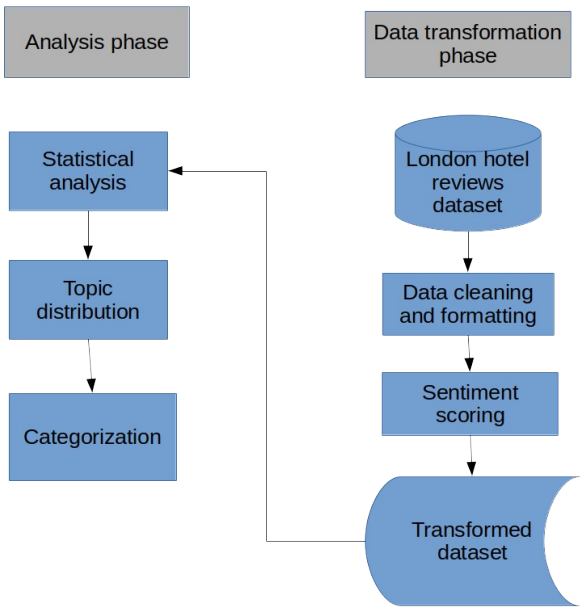


FIGURE 2

DETAILED METHODOLOGY

Data cleaning

To begin, the csv file was loaded into Python with ISO-8859-1 decoding. All the columns were converted to lower case to ease the further processing of the data. After that, the whole text is tokenized using nltk (Natural language toolkit) [10]. Following the tokenization, the data is stripped of punctuation and custom stopwords. The cleaned data is then saved as a separate csv file.

Sentiment scores

SentiStrenght is a standalone tool for evaluating positive and negative sentiment in short informal texts such as these hotel reviews. The tool is free for academic research. Another csv is created with the tool and then imported to python again. Overall sentiment is calculated in Python as a correlation of positive and negative sentiments.

Statistical analysis

Statistical analysis began by calculating Pearson correlation coefficients for hotel review ratings and positive, negative and overall sentiments.

For each hotel, a mean of review ratings along with their positive and negative sentiments was calculated with python's *mean()* function. Second order statistic (standard deviation) was also calculated with the function *std()*.

If a hotel's user rating standard deviation was higher than our selected limit of 75 %, the hotel was marked as ambiguous. For each of the ambiguous hotels, a classification was done to either negative or positive subclass depending on which type of reviews the hotel had more.

A good and commonly used tool - Python's matplotlib - was utilized to visualize our findings. First, a histogram of hotels with high standard deviation was plotted for each star category. Also, a histogram was drawn of the proportion of the two subclasses. To further visualize the content of the subclasses, wordclouds were drawn with Python's wordcloud module for the both subclasses. These wordclouds highlight the most frequent words used.

Topic distribution

Before creating positive and negative LDA models, the words needed to be preprocessed with SnowballStemmer, gensim and other modules. LDA was used with five topics and three words per topic to determine the topic distribution of the subclasses.

Empath Client is another Python tool to analyze text across lexical categories and has some built-in categories for this. The LDA models were analyzed with Empath to count the numbers of occurrences in each category. Next, the number of found categories was divided by all the Empath categories to get an indication of agreement between Empath and common corpus. The agreement was also calculated between Empath categories and LDA.

Categorization

The corpus was gathered into six standard categories: Price, service, parking, room, location and food. For this, Python nltk wordnet module was used. The words were assigned to either positive or negative set and into a standard subcategory within them using Wu-Palmer Similarity algorithm.

After categorization, the frequency of each subcategory in positive and negative categories was plotted in bar graphs.

RESULTS AND DISCUSSION

The sanitation of our data had left us with 27010 rows of data, which equals to same amount of reviews. Unfortunately, we were not able to remove every single foreign word from this data. The sanitation removed most of the foreign words, but some words remained since our technique of removing the words was limited. Our implementation checks whether a word exists in the English dictionary but unfortunately some overlapping was discovered later in the project. This remains as one limitation of our project but luckily it counts as a very marginal error, since the valid data amount is still very large compared to the invalid words.

The initial sentiment analysis resulted in 20471 positive sentiments, 1793 negative sentiments and 4746 neutral sentiments. From these, the person coefficient was calculated and then the second order statistics were calculated from the hotels and high variance hotels were separated for further analysis from a chosen threshold of

0.65. This gave us six different hotels to analyze the reviews from.

The hotels that were above the threshold were labeled as ambiguous. The amount of ambiguous class positive reviews ended up being 3978 and the amount of ambiguous class negative reviews was 640 while neutral reviews ended up being 1549.

Figure 2 shows the word cloud for the negative subclass of these ambiguous hotel's reviews. Some overlapping was noticed on the most common words like breakfast, good, and night. But the words after that differ some to the positive word cloud which can be seen in Figure 3.

FIGURE 2

NEGATIVE SUBCLASS WORD CLOUD

FIGURE 3

POSITIVE SUBCLASS WORD CLOUD

From here we can already tell some differences that cause the negative reviews and positive reviews. For example, negative reviews have quite large frequency of the word "dirty", also things related to hygiene, such as *dirty*, *bathroom*, *clean* and *toilet*. Clean in the negative hotel reviews often are accompanied by not, so this is why it overlaps with the positive reviews, but it semantically it has ended up in the negative subclass. This is one clear benefit of doing semantical analysis and not just statistical calculation of the words. The noteworthy mentions of the positive word cloud includes *service*, *staff*, *clean*, *best* and *definitively*. One of the most important achievements of a hotel is to get a renewing customer so the positive sentiment of the word *definitively* is worth analyzing to see which cause is making the customer to "*definitively*" come again.

We analyzed the hotel's star reviews to see which hotels are more high-end and which are low-end. These ratings can be seen from the histograms made from each star rating categories. Figures 4 to 8.

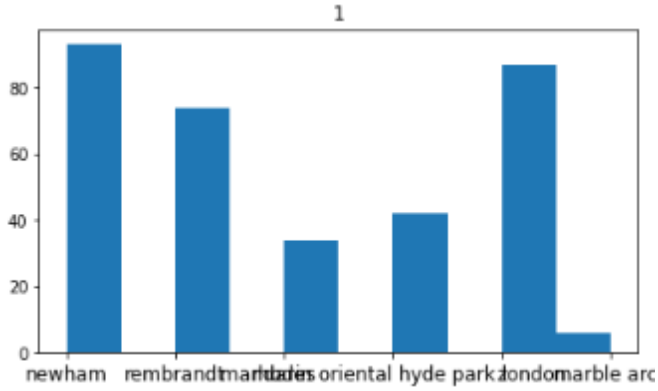


FIGURE 4

ONE STAR RATING FREQUENCY

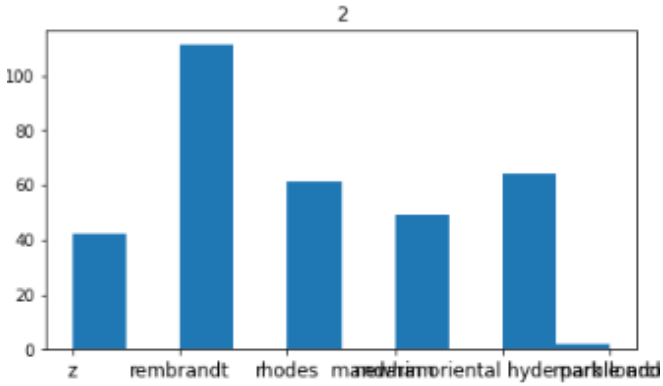


FIGURE 5

TWO STAR RATING FREQUENCY

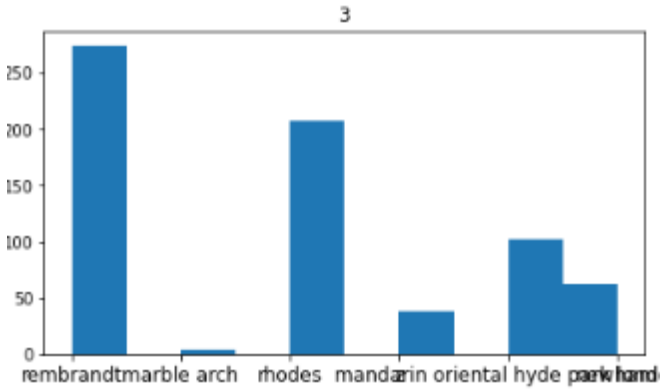


FIGURE 6

THREE STAR RATING FREQUENCY

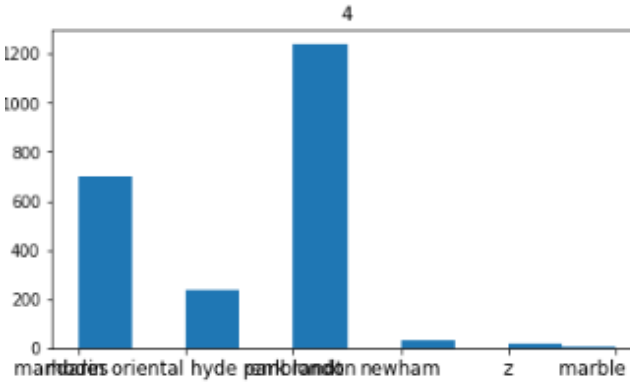


FIGURE 7

FOUR STAR RATING FREQUENCY

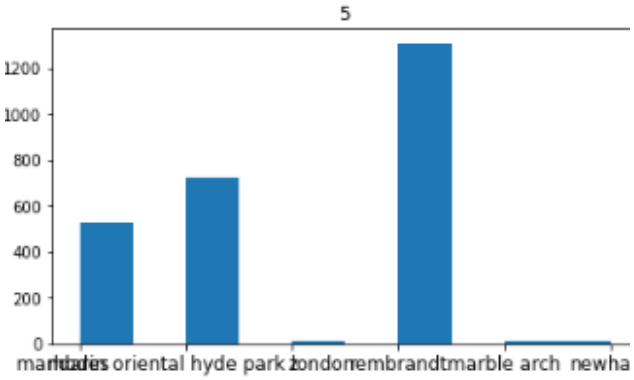


FIGURE 8

FIVE STAR RATING FREQUENCY

The LDA performed to the negative and positive subclasses produces topics of the reviews. These resulted in topics shown on Figure 9 and 10.

Topic	Word1	Word2	Word3
0	staff	clean	servi
1	camera	park	locat
2	locat	friend	night
3	walk	great	small
4	breakfast	good	staff

FIGURE 9

POSITIVE LDA TOPICS

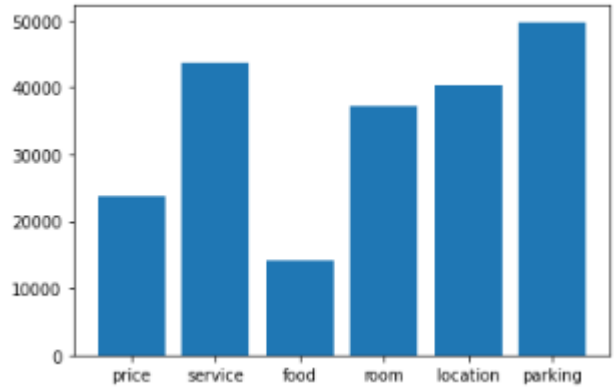
Figure 9 shows the positive topics created by the LDA model. We did not name these topics but one could possibly use to overlapping topics from the empathy analysis to put a topic name for the LDA models that would fit.

Topic	Word1	Word2	Word3
0	breakfast	bathroom	shower
1	camerafare	bathroom	solo
2	night	place	book
3	shower	bathroom	work
4	staff	clean	dirti

- FIGURE 10

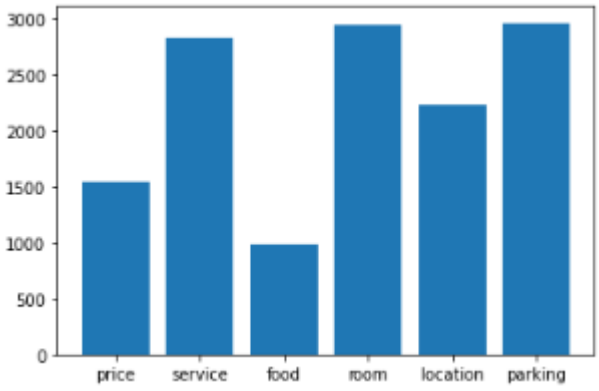
- **NEGATIVE LDA TOPICS**

- The negative LDA topics can be seen from the Figure 10. The words of these topics are stemmed which is good to keep in mind when looking at them.



- FIGURE 11

- **POSITIVE SUBCLASS FREQUENCY OF COMMON HOTEL REVIEW TOPICS**



- FIGURE 12

- **NEGATIVE SUBCLASS FREQUENCY OF COMMON HOTEL REVIEW TOPICS**

Finally, the all the words of the reviews were placed under the closest category of the common hotel categories. This was done using the Wu-Palmer Similarity. The positive subclass category frequency can be seen from Figure 11. The negative subclass category frequency is shown on Figure 12. From this we learned that negative reviews are

most often related to the room, parking and service. On ther other hand, the positive frequency shows that the room had lower meaning than the negative as well as the service. Which means that in order to get rid of most of the negative reviews the hotel managers should more focus on improving the service and the room conditions. The room conditions mostly consider the hygiene and cleanliness according to the empathy classes and topics. The service on the other hand concerns mostly on politeness.

For future research we would suggest on trying to automate the combination of this type of data that was achieved in this project. The combined data of the LDA model topics and empathy categories would be a very fast way to analyze a large set of hotel reviews. This is most likely also applicable for other type of reviews also, but then of course, it would need to be fine tuned to that particular goal. This has great potential for a business opportunity to quickly seek for hotels that need improving and offering a solid plan for the hotel to work on.

Since most of the frameworks in state of the art leave this type of automation out, it could be also beneficial to see if a improved framework could be done with this method, by empirically verifying the value of this type of approach.

CONCLUSION AND PERSPECTIVES

Luckily, all of the group members had previous experience with Python, so the main focus could be on the tasks themselves rather than learning a new programming language. Hence, most of the tasks got done with only the task 12 missing due to its complexity. Rest of tasks were easy enough to do within the given time frame.

The group overall got much more fluent with natural language processing processes. It was nice to see that the theory of the lectures was well implemented in the exercise and a lot from the theory was able to take advantage when we had to figure out how we would overcome these tasks.

The group members worked individually on the tasks assigned to them but assistance was readily available from the rest of the group. Workflow was smooth and administrative tasks such as meetings took little time.

This project introduced us some useful tools for natural language processing and text analysis. To further process the data, a way of easily extracting and composing all the data for a single hotel could be developed. This would provide the hotel owners a way to utilize the data and develop their business.

REFERENCES

[1]Kasper, W., & Vela, M. (2011, October). Sentiment analysis for hotel reviews. In Computational linguistics-applications conference (Vol. 231527, pp. 45-52).

[2]K. Zvarevashe and O. O. Olugbara, "A framework for sentiment analysis with opinion mining of hotel reviews," 2018 Conference on Information Communications Technology and Society (ICTAS), Durban, 2018, pp. 1-4, doi: 10.1109/ICTAS.2018.8368746.

[3]V. B. Raut and D. D. Londhe, "Opinion Mining and Summarization of Hotel Reviews," 2014 International Conference on Computational Intelligence and Communication Networks, Bhopal, 2014, pp. 556-559, doi: 10.1109/CICN.2014.126.

[4] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1–135, 2008.

[5] B. Liu, “Opinion mining and sentiment analysis,” Handbook of Natural Language Processing, 2010.

[6] Kaggle - https://www.kaggle.com/PromptCloudHQ/reviews-of-londonbased-hotels

[7] SentiStrenght - http://sentistrength.wlv.ac.uk/

[8] Jupyter - https://jupyter.org/about

[9] Empath client - https://github.com/Ejhfast/empath-client

[10] Natural language toolkit - http://www.nltk.org/