1. **Introduction. An overview of the project and an outline of the shared work.**

The main topic of our project is analyzing Naïve Bayes, RBF kernel support vector machine and linear vector machine models with lasso, in order to get the best text classification. Our work on the project can be divide into four major steps: 1. exploratory data analysis 2. data cleaning and mining 3. Algorithms choosing and setting up 4. Conclusion and Summary

We didn't clearly divide these steps into certain person, but we mixed up the individual work into each of these steps and at same time, each person can have at least two or three focus area to work.

2. Description of your individual work. Provide some background information on the development of the algorithm and include necessary equations and figures.

I have worked on the most work of the exploratory data analysis, data mining, preparation and setting up for the RBF kernel model, and the conclusion for the possible improvement.
We used the below equations for linear SVM (learned from lecture):

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \qquad (1)$$

$$s.t.\ y_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1 - \xi_i \qquad i = 1, \ldots, m$$

$$\xi_i \geq 0$$

$$\min \|X\beta - b\|_2^2 \qquad (2)$$

$$s.t.\ \sum_{j=1}^{d}|\beta_j| \leq t$$

For RBF kernel, I use the RBF kernel equation:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right)$$

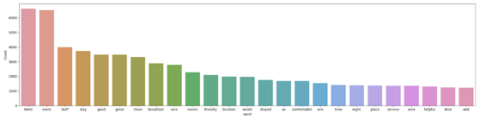
3. Describe the portion of the work that you did on the project in detail. It can be figures, codes, explanation, pre-processing, training, etc.

Firstly, I did the most of work of exploratory data analysis. The goals of this part is to have better understanding of the text data. By using the Pandas, I went through the several steps and methods, such us checking and summarizing the number of words, number of characters, Average word length and number of stop words etc. I found out some of the stops are actually useful for our choice to do the text mining preprocessing. For example, by doing the most common/rare words, we can tell that the step of removing stop words is needed, since most of the common words are the stops words. However, at same time, we can't delete them all, cause some of the very common words are highly related to the rating.

Secondly, after getting the useful information from the exploratory data analysis, I started to go through the text mining preprocessing, such as lower case, remove punctuation, removal of stop words and lemmatization etc. The codes method I majorly used is apply lambda x, which is the efficient way to apply the function to every row of the column.

Thirdly, I did the preparation of the Algorithm set-up. The process included train, test split and "Bag-of-Words" by using CountVectorizer ().

4. Results. Describe the results of your experiments, using figures and tables wherever possible. Include all results (including all figures and tables) in the main body of the report, not in appendices. Provide an explanation of each figure and table that you include. Your discussions in this section will be the most important part of the report.



These two graphs gave up the clear comparation before and after the mining processing. It shows the preprocessing methods, such as removing common words, stop words, rare words, lemmatization, stemming, and spelling correction reduced our vocabulary cluster so that features produce for the classification in the end are more effective.

Accuracy Rate of RBF without scale:0.462
Accuracy Rate of RBF with NMP Scale: 0.7436

About the discussion of the model, there are two result worth to mention:

The 46.20% is the accuracy rate of using the RBF Kernel to predict the people's rating on the hotels. It is actually not a very good prediction compare with the linear SVM.

The 74.36% is the new accuracy rate with rescaling the rating variable using negative (for ratings from 0 to 3), Moderate (for ratings from 3 to 4) and positive (for ratings from 4 to 5).

5. Summary and conclusions. Summarize the results you obtained, explain what you have learned, and suggest improvements that could be made in the future.

Comparing the accuracy rate of RBF and linear SVM, we will choose the linear vector machine. Here, I came out with the possible reasons. The text has a lot of features and the linear kernel is good when lots of features are used. It is doesn't really help to increase the dimensional space like using the RBF Kernel. Second, analyzing the text related information means that it take more time. Training with linear kernel is faster. Third, comparing with the other kernels, the linear need less parameters to optimize when train the data.
I also found out most of the comments are subjective that people may use their own "rating system" to give the their own hotel rating. Also, people are more likely to give the 4+ or 5 score even they have slightly positive hotel experience. It can be the limitation by nature on predicting using the "1 to 5" rating system. To get a better outcome, I went back to the preprocessing stage and did the test to use different scales on our hotel rating column. if I rescale to the negative (for ratings from 0 to 3), Moderate (for ratings from 3 to 4) and positive (for ratings from 4 to 5), our accuracy rate will significantly increase from 46.20% to 74.36%.

6. Calculate the percentage of the code that you found or copied from the internet. For example, if you used 50 lines of code from the internet and then you modified 10 of lines and added another 15 lines of your own code, the percentage will be 50 − 10 50 + 15 × 100.

(50-25)/ (50+45) = 26.32%

7. References.

A. N. (n.d.). *Support Vector Machines*. Lecture

http://cs229.stanford.edu/notes/cs229-notes3.pdf

Text Analytics for Beginners using NLTK. (n.d.). Retrieved from https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

J., S., T., & R. (n.d.). *1-norm Support Vector Machines*. Stanford University

https://papers.nips.cc/paper/2450-1-norm-support-vector-machines.pdf