Project Progress Report - Hotel Review Analysis

* Teammember:
  + Xin Qu: [xinq2@illinois.edu](mailto:xinq2@illinois.edu)
  + Biruo Zhao: [biruoz2@illinmois.edu](mailto:biruoz2@illinmois.edu) (Coordinator)
* What’s been done:
  + Data cleaning has been done, including removing duplicated data, checking missing values and refilling “NA” data. In original dataset, there are “nothing” and “everything” is both “Positive\_Reviews” and “Negative\_Reviewa”. In order to get all candidates review texts for future sentiment analysis, change “nothing” in “Positive\_Reviews” to negative reviews and “nothing” in “Negative\_Reviews” to positive reviews, keep “everything” in “Positive\_Reviewa” still as positive reviews and “everything” in “Negaitive\_Reviewa” still as negative reviews. In order to find the optimal machine learning approach, I set up binary categorizations for reviews, 0 stands for negative review and 1 stands for positive review.
  + Basic of several features of dataset has been done, e.g. word cloud for reviewers’ nationality, most common positive words, most common negative words, and bar chart of reviewed date and average score.
  + Data processing has been done,including tokenization, removing stop words and stemming.
  + Apply Naive Bayes, Logistic Regression, SVM and decision tree (from “sklearn” library) approaches to find the optimal approach by obtaining the highest accuracy score. Naive Bayes with bi-gram has the highest accuracy value.
* What remains to be done:

Will preform aspect sentiment analysis for most 5 common words in both positive reviews and negative reviews. Write a function to predict whether the provided review is positive or negative and write a function to predict the score by providing review text. Probably return a ranked list of hotels under each top 5 most common aspects in positive reviews and negative reviews. Try Stanford CoreNLP library to see how results vary.

* Barriers:

1. By applying removing stop words “from nltk.corpus import stopwords” and stemming by “from nltk.stem import PorterStemmer”, I found out “room” and “rooms” are treated as two different words and “us” is not in the “from nltk.corpus import stopwords”. --Solution: download a “stopword.txt” and apply removing stop words again to get rid of some common words, such as “us”. Compare most 5 common words by applying two ways of removing stop words, see if there is difference.
2. In sklearn library, how to apply most common word in both positive and negative reviews with various score? Such as , if X shows in positive reviews, the score of X is +??, if shows in negative reviews, the score of X is -??. How to set up score for each word?
3. From the column “Average\_score”, the average score is in range [5.2 , 9,8], most scores are around 8.0. It’s a little bit hard to find the threshold for positive review and negative review. --Solution: set any score >=7.0 as positive? Any other method?