Project Proposal—Hotel Score Prediction and Sentiment Analysis

- What is the function of the tool?
  This tool is designed to predict a score of a hotel in Europe and visualize the weakness based on its reviews. The dataset we use is hotel reviews data in Europe from *booking.com*.
- Who will benefit from such a tool?
  Hotel manager will benefit a lot from this tool. Hotel manager could predict the score from new reviews and figure out the weakness from past reviews for future improvement.
- Does this kind of tools already exist? If similar tools exist, how is your tool different from them? Would people care about the difference?
  Yes, similar tools exist. Based on the text analysis of positive reviews and negative reviews, this tool provides a possibility to predict whether the given review is negative or positive and determines how positive or negative the review is. Various classification methods will be applied and final classification method will be determined by the highest value of accuracy. Another function is to categorize the areas of most positive reviews and most negative reviews for future improvement. We plan to apply more than 2 categories, such as service, food, parking, etc., and perform sentiment analysis of each category. Tasks may change due to time limitation.
- What existing resources can you use?
  May use useful libraries, such as pandas, math, numpy, nltk, sklearn, matplotlib, etc., and will construct tool by hand.
- What techniques/algorithms will you use to develop the tool?
  - Data parsing / tokenizing input text, removing stop word, stemming
  - Different classifier methods (NaiveBayes, LinearSVC, KNN, …)
  - Classification evaluation: accuracy score for Sentiment analysis (may apply multiple tools including Stanford sentiment analysis module)
- How will you demonstrate the usefulness of your tool?
  We plan to split raw dataset into training data and test data, build average score model based on training data by obtaining the largest value of accuracy and calculate the accuracy for test data. At last, we plan to get as high accuracy value of test data as we can and visualize results.
- A very rough timeline to show when you expect to finish what. (total 8 weeks)
  - data processing and raw data analysis, including visualization—1 week
  - stemming, tokenization—0.5 week
  - feature extraction from reviews—1.5 weeks
  - classifier model and evaluation—3 weeks
  - final report—2 weeks
- Team member:
  - Xin Qu: xinq2@illinois.edu;
  - Biruo Zhao: biruoz2@illinois.edu – Project coordinator