Project Stage 1 - Report

Team: Akila Nagamani (nagamani@wisc.edu)
Arjun Kashyap (akashyap3@wisc.edu)
Meghana Moorthy Bhat (mbhat2@wisc.edu)

1. **Data Information and Links**

- **Link to the directory containing 300+ documents (RawData)**
  - **https://github.com/meghu2791/DataScience/tree/master/RawData**

- **README - Contains details about the entity type and markup tags used to mark the documents.**
  - **https://github.com/meghu2791/DataScience/blob/master/README**

- **Link to the test data set**
  - **https://github.com/meghu2791/DataScience/tree/master/DataSets/TestDataSet**

- **Link to the train data set**
  - **https://github.com/meghu2791/DataScience/tree/master/DataSets/TrainDataSet**

- **Link to the compressed folder**
  - **https://github.com/meghu2791/DataScience/blob/master/hotelNameClassifier.zip**

2. Stage 1 results and statistics

- **We have marked the hotel names based on data reviews from TripAdvisor.**
- **Our implemented extractor will classify any given input to hotel and non-hotel entities.**
- **We have marked 300 documents out of which 200 are used for train data sets (Set I) and 100 are test datasets (Set J).**
  - **We have ~12000 positive examples in train set (Set I) and around ~2000 positive examples in test set (Set J).**
  - **We have ~35000 negative examples in train set (Set I) and around ~5000 in test set (Set J).**
- **Classifier M - After first round of cross-validation, we found Decision Tree the better of other classifiers. The figures are below in the table:**

Classifier M (Decision Tree)	Percentage
Precision	45%
Recall	37%
F1 score	28%

- We did not do any post processing rules to further improve the F1 score. Instead we added few pre-processing rules and added new feature attributes to the model.
    - As example, we added stop words like prepositions in the English grammar to remove these keywords from the consideration in the code.
    - We added new features like ‘room’, ‘stay’ to the model which enhanced the performance of the classifier in the later turns.
- With these additions, we found SVM to outperform Decision Tree. Hence we chose SVM as our classifier for the test set J with the following statistics:

Classifier M (SVM)	Percentage
Precision	90%
Recall	83%
F1 score	86%