CS 838 (Spring 2017): Data Science
Project Report - Stage 2 (Group 12)

Deepanker Aggarwal
deepanker@cs.wisc.edu

Saket Saurabh
ssaurabh@cs.wisc.edu

Vishnu Lokhande
lokhande@cs.wisc.edu

1 Objective

In this stage of the project, we do information extraction from text documents using supervised machine learning approaches. We have multiple text documents and the goal is to extract specific kind of entity from these text documents.

2 Dataset preparation procedure

In this supervised classification approach, we manually labeled the required entities (the dish names, in our case) in all the text documents. We divide the data-set into two sub-datasets, the first one being a training (or development) set and the second one is a test set.

3 Dataset and Entity description

Each text document pertained to a restaurant review in the New York City. The development set had 215 documents and the test set had 100 documents. The entity type which we extracted are the names of the dishes mentioned in these text documents. We marked up the dish names in reviews using **<dish>...</dish>** tags. Some examples of the dish names found in the dataset are as follows:

- Caesar salad
- Spinach risotto
- Chinese soup dumplings
- Smoked fish platters
- Fried Chicken wings
- Papas Fritas
- French toast
- Samosas

A summary of dataset is given as follows.

Total number of text documents	315
Total number of mentions	3281
Number of documents in development set, Set I	215
Number of mentions in set I	2266
Number of documents in test set, Set J	100
Number of mentions in Set J	1015

- The **Classifier M** obtained in our case was **Random Forest**. We report its precision, recall and F1 obtained for the first time on Set I below:
  - Precision: 78%
  - Recall: 28%
  - F1 score: 0.41
- The final **Classifier X** obtained in our case was **ADA Boost**. We report its precision, recall and F1 obtained for the first time on Set J below:
  - Precision: 92%
  - Recall: 65%
  - F1 score: 0.76
- Since our precision was high enough (>= 90%), we did not add any post-processing rules to our information extraction process.
  Hence, the **classifier Y** and classifier X are the **same** in our case.

4 Classification results

For the information extraction process, we obtained the best accuracy using ADA Boost classifier with a precision of 92% and a recall of 65% on the test dataset.