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

Group ID: 17
Gautam Singh(gautamsingh@cs.wisc.edu)
Harsh Singhal (hsinghal@cs.wisc.edu)
Naman Agrawal (namanagr@cs.wisc.edu)

Objective : Information extraction from text documents using supervised learning techniques

In this project stage, we perform information extraction from natural text documents using supervised machine learning techniques and extract all mentions of a certain entity type (name of food/beverage items) from restaurant review text documents.

Entity type

We tagged mention of name of food items and beverages in 400 text documents containing restaurant reviews. We have marked up mention of food items in each document within and tags. Whereas, negative examples have been tagged within <n> and </n> tags. For example, mention of coffee in a restaurant review document have been tagged (** Coffee **) as positive example and negative examples like delicious have been tagged as **<n>delicious</n>** .

Data-set: Development and Test sets

As per project specification, we divided randomized data-set in two categories, development/training set(Set- I) and test-set (Set- J). We used set-I for training and set-J to report accuracy of our learning based extractor. Number of documents in data-sets and number of mentions of the entity type in each data-set is summarized below.

Data-Set	Number of text documents	Number Of mentions of the entity type	Number of negative examples tagged
Set-B(Set-I + Set- J)	400	1512	699
Set-I	200	882	379
Set-J	200	630	320

Classifiers

As per project specification, we used following classifiers from scikit-learn package.

- Decision Tree
- Random Forest
- Support Vector Machine
- Linear Regression
- Logistic Regression

We used cross-validation to select best classifier from those listed above based on Precision, Recall and F1 values. We select classifier with highest precision as classifier X which is used to calculate accuracy on set-J.

Classifier Type	Precision	Recall	F1
Decision Tree	87.0	88.0	87.0
Random Forest	88.0	88.0	87.0
SVM	89.0	89.0	89.0
Linear Regression	64.0	45.0	39.0
Logistic Regression	92.0	92.0	92.0

Logistic Regression is Classifier-X as we obtained highest precision with this classifier.

We applied classifier-X on Set-J to check accuracy of our extractor and got result tabulated below.

Classifier Type	Precision	Recall	F1
Logistic Regression	94.0	94.0	94.0

Results

We selected Logistic Regression based classifier as our best classifier (Classifier Y) based on Precision, Recall and F1 results obtained from experiments. Results obtained with classifier Y on Set-J is tabulated below. We did not apply any rule based post-processing as we achieved expected precision with classifier-X. In other words, Classifier-Y is same as Classifier-X .

Classifier Type	Precision	Recall	F1
Logistic Regression	94.0	94.0	94.0