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 approach and extract all mentions of a certain entity type (name of food/beverage items) from food review text documents.

Entity type

We tagged mention of name of food items and beverages in 300 text documents containing food reviews. We have marked up mention of food items in each document within <p> and </p> tags. Whereas, negative examples have been tagged within <n> and </n> tags.

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	300		
Set-I	200		
Set-J	100		

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			
Random Forest			
SVM			
Linear Regression			
Logistic Regression			

cuz is Classifier-X as we obtained highest precision with this classifier. We applied classifier-X on Set-J to

Results

We selected   xyz 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 listed below.

Type of Classifier Y	Precision	Recall	F1