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

4 Feature Design and Extraction

For the purpose of information extraction, we created 17 features for each positive/negative example. Our feature design process was guided by the fact that common mention of dish names in each of these reviews had certain similar characteristics, e.g., often mention of a dish name was accompanied by the mention of ingredients for that dish within the same sentence. This motivated us to extract the sentence in which the example was present and we call the extracted sentence as the **context** of the positive/negative example. This context allowed us to come up with interesting features like, whether there was a mention of a 'food adjective' (an adjective often qualifying food like- sweet, salty, crispy, tasty, etc.), whether there was a mention of price in the same sentence, etc.

The following lists the 17 features that we used in our classification process:

1. count of food adjectives in dish name (dictionary based detection)
2. count of food ingredients in dish name (dictionary based detection)
3. count of food adjectives in context (dictionary based detection)
4. count of food ingredients in context (dictionary based detection)
5. has price mention (this is detected by the presence of dollar sign)
6. price distance (how far is the dollar sign from the actual example)
7. has meal name mentioned (common nouns like lunch, breakfast, etc.)
8. has dish quantity mention (common quantifiers like a cup of, a bowl of, etc.)
9. meal name in context
10. dish quantity in context
11. country context (dictionary based detection)
12. how far is a comma from the dish name
13. how far is a semi colon from the dish name

14. dish name starts with capital
15. count of capital letters in the dish name
16. count of commas in the dish name
17. count of semicolons in the dish name

5 Classification

We evaluated a number of classifiers from the sci-kit learn package for our purpose- SVM, logistic regression, linear regression, neural network, decision tree, random forest, and ADA Boost. We used five-fold cross-validation to choose the best classifier.

- 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 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.

6 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.