University of Wisconsin-Madison

CS 838 Data Science

Project Stage 2

Zomato vs Yelp

Team:
Sanjay S Seshadri
Sabareesh Ganapathy
Pallavi Ghosh

Instructor:
Dr. An Hai Doan

March 1, 2017

Overview of Project Stage 2

Following are the details and brief descriptions of the process steps in Stage 2 of the Data Science project. The aim of this project stage was to extract all mentions of a certain entity type (in our case, adjectives describing food items) from the 300 documents that were collected as part of Stage 1 of the project.

1 Documents: Description

We collected around 300 documents in the form of reviews that were provided for the restaurants listed in the websites Yelp and Zomato. The documents have been collected from the reviews from the Yelp website.
Following is the link to all the documents: `https://drive.google.com/drive/folders/0B_ianmgUSV9rY05vUWtFZ21sWEk`

2 Entities: Description

We have chosen the adjectives that describe the food and the dining experiences of the restaurant goers for the restaurants listed in Yelp and Zomato websites. Some examples of adjectives:

- "neat" place just right of broadway theatres
- bringing "awesome" thai food to Austin
- "yummy" cocktails

3 Entities : Markup

We have marked up the adjectives manually in the review documents using the following tags : <adj >and </adj>. The adjective is sandwiched between the two tags for the markup.
Example:
yummy *cocktails* transforms to ***<adj >yummy </adj>*** *cocktails*

4 Development and Test sets

We divided our collection of review documents into two sets, set I and set J. The set I has been used for development (dev set), and the set J has been used for reporting the accuracy of our extractor (the test set).

- **Set I**: The set contains 220 documents with approximately 4145 mentions of the positive examples of the adjectives. It also contains approximately 6502 generated negative examples. All the marked documents are available under the Dev set folder at: `https://drive.google.com/open?id=0B0HMBhaIv6CrNjN6M3hkX2ZpNTQ`
- **Set J**: The set contains 100 documents, with approximately 1618 mentions of the positive examples of the adjectives. All the marked documents are available under the Test set folder at: `https://drive.google.com/open?id=0B0HMBhaIv6CrOUkxaWUxYTRRbGc`

5 Features

The features that have been used to detect the occurrence of entities are as follows:

- length in chars
- preceded by was / is / an / are / so
- preceded by another adjective
- preceded by very
- *is succeeded by noun*

6 Cross Validation

With the initial 10 fold cross validation with SVM, we gained a precision of 82% and recall = 3%. This was due to existence of many false negatives in our development set. Upon inspection, we also found that our features were not distinguishing enough.
As a solution we added a new feature, "is succeeded by noun" where the

noun could be the name of the food, place etc. Post this step, 10-fold cross validation with SVM gained precision of 84% and recall of 55%.
Also, we further went ahead and separated the feature *preceded by was / is / an / are / so* into separate vectors with a boolean value each rather than having a single boolean variable for all of them.
After this step, the 10 fold Cross Validation with SVM gained precision of 88% and recall of 60%

7 Classifiers

Following were the classifiers that we used from the Sci Kit Learn package

- Decision tree
- Random forest
- Support vector machine
- Linear regression
- Logistic regression

Classifier	Tuning-Technique	Precision	Recall	F1 Measure
Decision Tree	max_depth variation, optimum=8	88.82	69.55	77.23
Random forest	max_depth, min_sample_split, min_sample_leaf	90.51	70.15	78.75
SVM	Threshold variation	90.34	61.93	72.47
Linear regression	Threshold value = 0.703	90.05	52.3	65.90
Logistic regression	Threshold value = 0.96	90.22	51.50	65.1

Table 1: Classifier comparison table with 10 fold CV

8 Test Set Performance

We choose RandomForest as the best Classifier since it has high recall value for precision greater than 90 percent. Following values were obtained on test set:
Precision - 92.8
Recall - 70.47
F1 Measure - 80.11