Task 6 Hygiene Prediction Report

Objective

Predicting whether a set of restaurants will pass the public health inspection tests given the corresponding Yelp text reviews along with some additional information such as the locations and cuisines offered in these restaurants. Making a prediction about an unobserved attribute using data mining techniques represents a wide range of important applications of data mining.

Procedure

Method 1 — array representation from nltk words

At the very beginning, I tried to process all the training reviews(573 entires) using nltk, like what we did in task 2, removing stop words, stemming, remove low frequency words, etc. Then I collected the most frequent words (freq >= 3) from these processed training reviews and convert the training reviews to array representation list according to these most frequent words (1 means specific word exits in review, 0 means not). Likewise, I did the same thing for the rest 12573 reviews and convert them to array representations (e.g. [1,0,0,0,1,1,1, .]) .

Then I use SVC model in scikit-learn to train and predict labels.

Method 2 — array representation from refined nltk words

The result by method 1 is not satisfied since there are some high frequency words which are numbers or meaningless such as “on”, “at”, “\” etc. Therefore, I filtered out all the numbers and some strange short words from nltk processed reviews. Then convert both training and testing reviews to array representation list.

This time the result is a little bit better.

Method 3 — array representation by word occurrence

I converted processed reviews by occurrence according to frequent words, thus the array list becomes: [2,5,1,0,4, …]

This gives me the optimal result.

Method 4 — array representation from refined nltk words and average ratings

As introduced in instruction, there are additional information such as review count and average ratings of each restaurant. Thus I considered average rating as an entry in array representation list, the list looks like this: [2,5,1,0,4, … , 3.75]

The result shows a little bit regression from method 3.

Method 5 — array representation by review count and average ratings

This time I used review count as well, the array representation list is: [2,5,1,0,4,…, 20, 3.75]

Method 6 — array representation by count from refined nltk words, average ratings and SegPhrase result

After processed by nltk, I trained reviews using SegPhrase and got frequent phrases. Then I combine top 3000 frequent words, top 1000 frequent phrases and review_count and average rating as the reference.

Thus the representation contains three parts, first :frequent words, second: frequent phrases, third: review count and average rating.

But the result is even worse.

Analysis

The result with highest score comes from method 3.

I deleted frequent strange words and numbers from nltk processed reviews, in this way I could eliminate the possibility that some frequent outlier words affects the overall result. It's possible that someone prefer to use numbers or strange words in reviews, and give many reviews to different restaurants, thus affect other reviews.

Moreover, I think list with occurrence count is more useful in that it keeps the popularity of specific words in the review which could been lost by simply 0 and 1 list.

However, additional info such as review count and average rating does not help in improving predict results. Also the SegPhrase result doesn't help. In addition, I tested with several models such as Naive Bayes and LDA model to predict the hygiene condition, but the results are not as good as simply refined frequent words result.

So far I could only get 10 out of 15 points, and I haven't found a better way to improve.