CS579 Project Report

Restaurant Grading And Recommendation

-- Based On The Sentiment Analysis In Twitter

Team member:

YIZHI GU CWID: A20334298

JIE HUANG CWID: A20279698

1. Introduction

It’s always hard to decide ‘what to eat’ when we have a plenty of choices. Nowadays, people like post their good or bad experiences on Twitter when they go to a restaurant. So, we can recommend people the best one among several choices by analysis these reviews.

We use the data from the Yelp to build the training set. Because we can get the ratings of different reviews and the reviews themselves, we collect these data and training them. After we got the different values of different terms, we use the training data to test the tweets. We collect same amount tweets about several restaurants by using the Twitter API and put them into different files. By using the training data to value these tweets, we can find out the positive reviews and negative reviews. Counting these positive reviews and compares different restaurants, we can easily find out the best choice. After we done these, we can recommend the restaurant to people. Meanwhile, we also build another feature. We pick a random user who give a positive tweet to the best restaurant, and recommend the best choice to his friends.

As we already know, the Twitter is a big database, and we can easily find the friends of people. And Yelp is another big database, which can easily find the ratings of different restaurants. So we consider if we can recommend the restaurant to the twitter users, by analysis the tweets about the restaurant. And our hypothesis is that the people’s tweets can help us rating the different restaurant.

2. Data

For the data collection, we have two steps.

First we collect the data from Yelp, we download the ‘yelp_academic_dataset.json’ file. From the file, we take 1000 positive reviews and 1000 negative reviews to be the training files. And we take another 1000 positive and 1000 negative reviews to be the testing files. After the training, we use the testing files to test the accuracy.

Second we request the tweets by using Twitter API. We set a restaurant file 3 restaurants as sample example. And request 100 recent tweets of different restaurants and put them into different files.

3. Methods

We use the machine learning method.

3.1 Use ‘tokenize_with_not’ to tokenize the Yelp training reviews. And vectorize and shuffle the data. By compare different parameter, we choose the higher accuracy. And get the training matrix.

3.2 Use the training matrix to predict the tweet is positive or negative. And calculate the number of positive tweets to choose the best restaurant of the restaurant list.

3.3 After we get the best restaurant, we use Twitter API to get the friends list of one random user, who has positive review of the top restaurant list.

4. Experiments

4.1 Do cross validation:

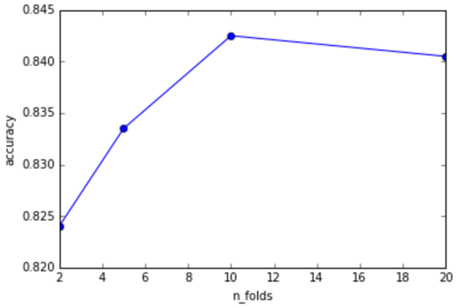
We pick the n_folds = 5 as we did in our homework.

```
fold 0 accuracy=0.85
fold 1 accuracy=0.82
fold 2 accuracy=0.8425
fold 3 accuracy=0.815
fold 4 accuracy=0.84
average cross validation accuracy=0.8335
```

4.2 Run one experiment, which consists of vectorizing each file, performing cross-validation, and returning the average accuracy. (Set min_df=1, max_df=1., binary=True, n_folds=5)

We get the accuracy using default settings: 0.8335

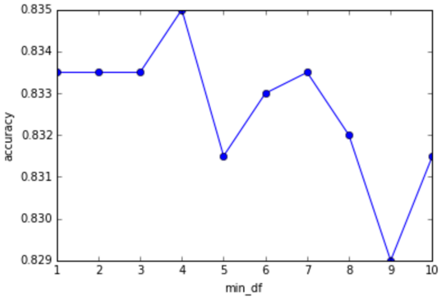
4.3Vary the setting of n_folds parameter (n = [2, 5, 10, 20])



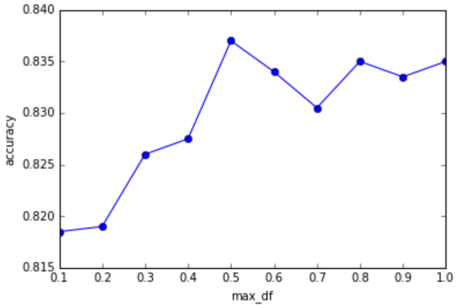
4.3 Setting binary twice, once with binary=True, and once with binary=False. And get higher accuracy when binary = True.

```
[0.83350000000000013, 0.82400000000000007]
```

4.4 Vary the setting of min_df parameter in the range (1,10). And we get top accuracy at min_df=4.



4.5 Vary the setting of max_df parameter to be [.1, .2, .3, .4, .5, .6, .7, .8, .9, 1.]. And we get top accuracy at mac_df = 0.5



4.6 Final Result

So, based on the above experiments, we set:

binary=True，min_df=4，max_df=.5 This results in 5-fold cross-validation accuracy of 83.7%.

And after we test another 2000 files our testing accuracy is 85.9%, which is pretty close to our estimated accuracy of 83.7%.

5. Related Work

Our recommendation system is related to other sentiment analysis. But our recommendation system is not just a simple sentiment analysis system. In our system we try to recommend one's favorite restaurant to his friends first, because twitter can supply relationship-net between twitter users. If you want to hang out with friends, our system can give you a better option.

6. Conclusions and Future Work

Limited by data scale, we can't perform too much analysis on twitter users. If possible, I really hope to build a system with all twitter data that can perform a Cascades and Clusters analysis on a specific user's relation to give a popular restaurant in their circle.

Reference

[1]. Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the 20th international conference on Computational Linguistics

[2]. sklearn.feature_extraction.text.TfidfVectorizer
http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfV ectorizer

[3]. sklearn.linear_model.LogisticRegression
http://scikit-learn.org/stable/modules/generated/
sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression

[4]. sklearn.cross_validation.KFold
http://scikit-learn.org/stable/modules/generated/ sklearn.cross_validation.KFold.html#sklearn.cross_validation.KFold