CS412 Project Report

What's Cooking?

Team: AllEatable

1. Team Members

He Huang	Ye Liu	Zhu Wang
Congying Xia	Lichao Sun	Fan Zhu

2. What we have done

Our problem is to classify where does the food come from according it's ingredients. We used 5 different machine learning methods, including Neural Network, LDA, Naïve Bayes, Random Forest, SVM, to build classifiers. Here's is the progress of each method.

- Neural Network (Accuracy: 79.2%)
we use a neural network with three fully connected hidden layer, and each hidden layer is followed by a dropout layer. A softmax layer is appended as the output layer to generate the prediction probabilities. We use the batched gradient descent with momentum to learn the network parameters. The sizes of the fully connected layers are 1000-500-100. The learning rate is 0.01, with a decay of 10e-6, and the momentum is 0.9.

- Naive Bayes (Accuracy: 72.94%)
We use MultinomialNB from sklearn.naive_bayes to build Naive Bayes classifiers. We set the additive (Laplace/Lidstone) smoothing parameter alpha to 1.0, and use 5-fold cross validation. We also tried TF-IDF feature, BernoulliNB and sklearn.naive_bayes.

- Random Forest (Accuracy: 72.4%)
We use RandomForestClassifier kit in python sklearn to build random forest classifiers.

- SVM (Accuray: 76.6%)
We used Linear SVM model in python sklearn to build classifier.

- LDA
we have already got the probability for document belong to each topics and the main topic words for each topic, which contains the highest percent relevance to the related topic.

- Visualization
We use word cloud and PCA to visualize the training data.

3. What to be done

One on hand, we will try to improve our accuracy in each method.

- Neural Network
We will try to modify the network, tune the parameters and adding bath normalization to get higher accuracy.

- Naive Bayes
Tune the parameters of MultinomialNB to get higher accuracy.

- Random Forest
We will make feature selection and adjust parameters in Random forest to improve the accuracy.

- LDA
We will assign class names to topics and predict the test file.

- SVM
We will try to use SVM with soft margin or Kernel functions in this problem, since the real word applications are usually not linear separable.

On the other hand, we will consider Model Combination. In each method, we can use ensemble methods like bagging and boosting. Also, we can use the voting result of different methods as the final result. In the end, we're try to use visualization to help us present our result.

4. What has changed

We are not using the class-based association rule mining algorithm, since it is outdated and not popular.