CS 145 Project 1: Use recipe ingredients to categorize the cuisine

1. Project Description

Here is an ongoing Kaggle competition which ask you to categorize the cuisine based on the recipe ingredients. Some of our strongest geographic and cultural associations are tied to a region's local foods. This project is trying to predict the category of a dish's cuisine given a list of its ingredients.

In the training and testing dataset, we include the recipe id, the type of cuisine, and the list of ingredients of each recipe (of variable length). The data is stored in JSON format.

An example of a recipe node in train.json:

```
{
"id": 24717,
"cuisine": "indian",
"ingredients": [
    "tumeric",
    "vegetable stock",
    "tomatoes",
    "garam masala",
    "naan",
    "red lentils",
    "red chili peppers",
    "onions",
    "spinach",
    "sweet potatoes"
]
},
```

In the test file test.json, the format of a recipe is the same as train.json, only the cuisine type is removed, as it is the target variable you are going to predict. All data is available on CCLE under Week 6. You can download the same data from Kaggle too [1].

Submissions are evaluated on the categorization accuracy (the percent of dishes that you correctly classify).

2. File descriptions

train.json - the training set containing recipes id, type of cuisine, and list of ingredients

test.json - the test set containing recipes id, and list of ingredients

sample_submission.csv - a sample submission file in the correct format

3. Hints

This is a multiclass classification problem. You can use the classification methods covered in our lectures or other books. In its most basic form, this problem decomposes

trivially into a set of binary classification problems, which can be solved using the techniques we talked about. You can choose One-vs-All Classification or All-vs-All Classification [2] to decompose a multiclass classification problem.
You can also use bag of words model [3] to represent the recipe ingredients to a feature vector. For example, if the total number of ingredients is N, then you can use a N dimensional vector to represent a recipe, where each entry of the vectors refers to the corresponding ingredient.

4. Project Requirements

Please submit your prediction result on the test data to “make a submission” on the dashboard of this project on Kaggle. The username of your submission should be the same as your team name. Each team also needs to submit a project report on CCLE. In the first page of the project report, the team name, student names, UIDs, and the role of each team member are expected. There is no restrictions on programming languages. You are also allowed to use online packages. But you need to clarify that in your report. You also need to submit your code on CCLE and make sure your prediction results is reproducible. If you use randomized algorithms, similar prediction accuracy is expected. A good report should include baseline algorithms and show as many comparisons as possible. You are also expected to explain the comparisons. (The explanation, like “it gives high prediction accuracy”, is too general). Please try your best to understand the rationale behind each algorithm and use your own words to explain. Keep in mind that the project is evaluated not only on the accuracy performance but also heavily the clarity of the report. The report should be around 10 pages.

5. References

[1] Kaggle “What’s Cooking” https://www.kaggle.com/c/whats-cooking/
[2] Multiclass classification http://www.mit.edu/~9.520/spring09/Classes/multiclass.pdf
[3] Bag of Words model https://en.wikipedia.org/wiki/Bag-of-words_model