COMP 4651 Cloud Computing and Big Data Systems – Project Proposal

Analysis of the Ingredients in the Cuisine Around the World

LIU, Qinhan (20328923, Handle: qliuan), WU, Wei (20328642, Handle: wwual),
YANG, Xin (20328276, Handle: xyangaq), YE, Ziyun (20329074, Handle: zyeai)

Background and Introduction

Many people enjoy traveling around the world and tasting foods in different cultures is a very important and unavoidable part of that. Sometimes, it may be a struggle for people to make a decision for choosing a dish that they are unfamiliar, especially when they are overwhelmed with choices. Even though there are some traveling advising and restaurant reviewing APPs, travelers may have their own preferences of different cuisine. Since in most menus, ingredients of dishes are usually provided, it would be much helpful for travelers to make decisions if they can guess which cuisine the dish may belong to based on the ingredient. Therefore, in the hope of helping travelers to order their food, here we want to find the most commonly used ingredients within or across various cuisines in order to provide guidance for travelers. We would like to work on a dataset on Kaggle called 'What's Cooking?' which collects the ingredients of more than 3,000 recipes from 20 different cuisines around the world.

Proposed analysis tasks

To solve the problem we address, four analysis tasks are proposed to be performed on the dataset:

1) By calculating the frequencies of the ingredients used in different cuisines, a word cloud, which reflects the general impression of a certain type of cuisine, can be generated.
2) Another word cloud, which indicates some ingredients that are commonly used around the world, can be obtained by calculating the frequencies of the ingredients used in all cuisines.
3) Different types of cuisines can be clustered with user-defined measurement of their similarities in their ingredient frequencies. Then, the similarities between cuisines can be compared in the visualization.
4) With the patterns of the current dataset we obtain, we will try to classify the un-labeled recipes with their ingredients using machine learning methods like Bayes' Net. The results can be evaluated by comparing to the patterns of labeled recipes.

System Overview and Potential Technical Challenges

In our project, there is no need to worry about retrieving the data since the dataset is already provided in Kaggle. We will be using Spark and Hadoop on AWS EC2 for processing of the raw data and derivation of the frequency features. This should not be so difficult given that we have studied the program of counting word frequencies in class. To visualize the features, we may need matplotlib and d3.js, which provide versatile and convenient built-in functionalities for graphics. We do need to spend some time in getting familiar with the visualization tools since most of our group members do not have related experiences. For the clustering and classifying tasks, we may implement the algorithm on EC2 using Spark or stick to scikit-learn, the mature python machine learning package. We should be able to implement some reasonable algorithm using Spark or scikit-learn if we spend some time studying the programming APIs.