**COMP4651 – Cloud Computing and Big Data Systems**

**Project Report**

**Group Members:**

Zeng Yueran 20591873 yueransutd

Naik Hiong Chiang 20591885 frankienhc

Tan Ting Yu 20591861 tytanab

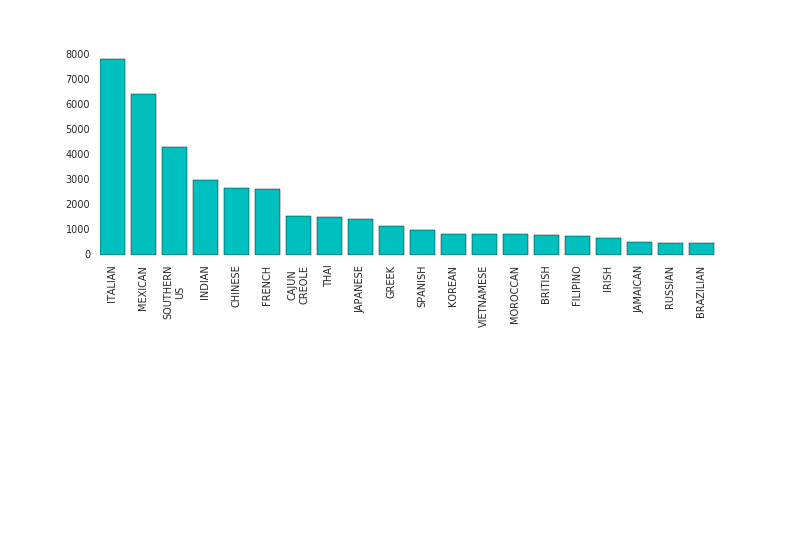
Introduction

Data Exploration & Processing

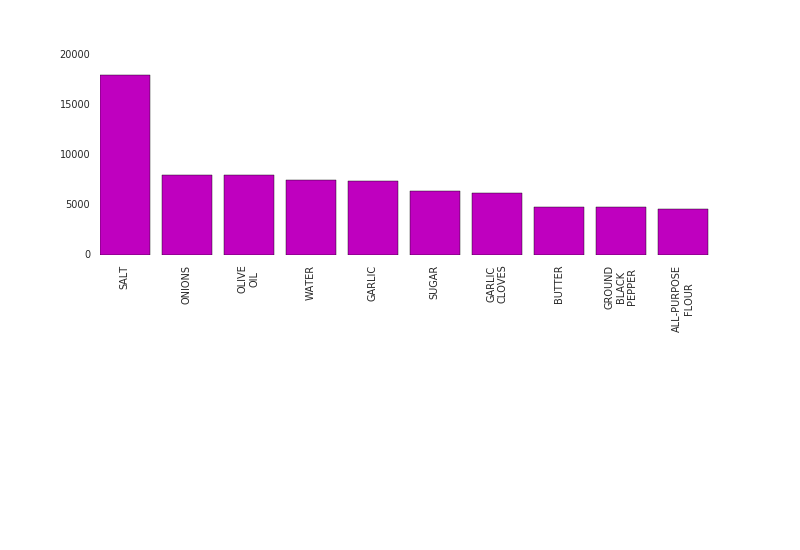
Before formulating our evaluation strategy, it is necessary to examine the given dataset, in order understand its limitations and ensure that is sufficiently well-balanced. There are a total of 39, 774 recipes across 20 cuisines, containing 6714 unique ingredients, with each recipe stored as JSON objects in the form:

**{"cuisine": "indian", "ingredients": ["tumeric", "vegetable stock", … ]}**

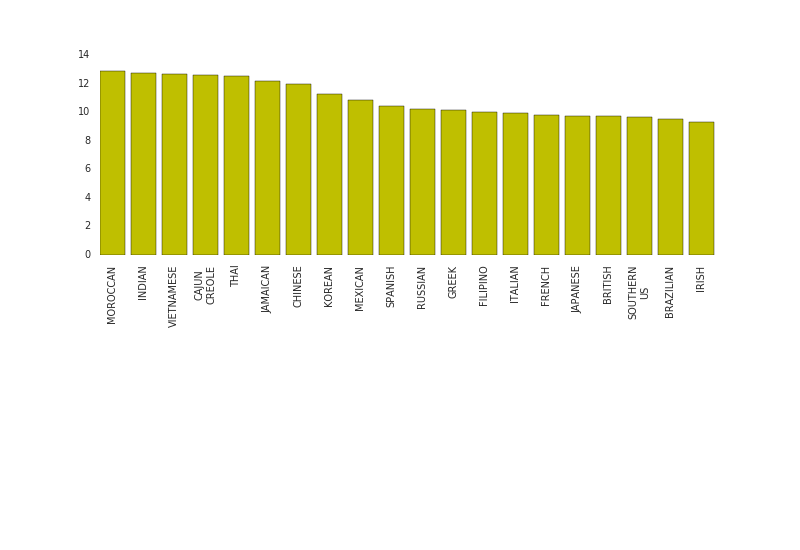
Applying MapReduce transformations with Spark, helpful visualizations that describe the characteristics and distribution of data can be constructed, as shown below:



*No. of Recipes By Cusine*

**

*No. of Occurrences Per Ingredient Across All Recipes (Top 10)*



*Average No. of Ingredients Per Recipe, By Cuisine*

Key observations include the following:

1. Italian and Mexican recipes are overly-represented.
2. Single-occurrence ingredients are too specific, e.g. ‘Kraft Original Barbecue Sauce’, ‘Boned Skinned Duck Breast Halves’, ‘Pineapple Chunks In Natural Juice’, thus not particularly helpful.
3. The average number of ingredients used per recipe is reasonably uniform across cuisines (= 10.914 ≈ 11).

Using Spark’s Word2Vec machine learning library, we can search for and sieve out ingredients that are essentially or close to identical. To demonstrate, we compare 5 ingredients that appear only once with all other ingredients in the dataset, so that, if possible, we can rename them accordingly and hence reduce the number of ‘features’.

|  |  |  |
| --- | --- | --- |
| **Ingredient** | **Most Alike** | **Cosine Similarity** |
| TOPSIDE STEAK | STEAK | 0.824 |
| CALAMARI STEAK | 0.749 |
| RUMP STEAK | 0.705 |
| ICING MIX | SOUR MIX | 0.857 |
| STUFFING MIX | 0.827 |
| CAKE MIX | 0.817 |
| KRAFT ORIGINAL BARBECUE SAUCE | BARBECUE SAUCE | 0.928 |
| HOISIN SAUCE | 0.915 |
| ASIAN BARBECUE SAUCE | 0.914 |
| SKINLESS CHICKEN BREAST FILLETS | SKINLESS CHICKEN FILETS | 0.992 |
| CHICKEN BREAST FILETS | 0.989 |
| BONELESS CHICKEN BREAST | 0.987 |
| BONED SKINNED DUCK BREAST HALVES | SKINNED BONED DUCK BREAST HALVES | 1.0 |
| BONED DUCK BREAST HALVES | 0.987 |
| DUCK BREAST HALVES | 0.975 |

Prediction Metric

To start, the data is separated into ‘training’ and validation sets using Spark’s randomSplit method, with a distribution of 99% and 1% respectively. This is practical as we are using a simple linear regression model – we will not be updating the weights once they have been calculated, thus we want to maximize the number of ‘training’ data points, which in this situation includes a comprehensive ingredients database.