We implemented Naive Bayes to predict the cuisines, because it is fast to process a very large data set. However, we realized that the probabilities could get very small and when they are multiplied together, the product would be close to zero and could cause underflows. In order to prevent this problem, we edited the code to take the logarithm of the probability of the ingredient given the cuisine for calculating the probability of the cuisine. Furthermore, if an ingredient does not exist in the given cuisine, we set the probability to 0.00000000001 instead of zero to prevent ignoring the probabilities of all other ingredients, because if the probability of one ingredient is zero then the probability of the entire set of ingredients given a particular cuisine would be zero.

Then we used k-fold cross validation to estimate the performance of the hypotheses that we obtained on unseen data. We shuffled the training set and divided it into 6 disjoint sets.

The time required for training was 0.533 seconds and the processor is 2.9 GHz and memory is 16 GB.

|  |  |
| --- | --- |
| Run # | Accuracy (%) |
| 1 | 75.25083612040133 |
| 2 | **71.90635451505017 (min)** |
| 3 | **71.90635451505017 (min)** |
| 4 | 74.24749163879599 |
| 5 | **76.5886287625418 (max)** |
| 6 | **76.5886287625418 (max)** |
| **Average** | **74.41471571906355** |