|  |
| --- |
| Recipe2Cuisine |

You've just joined the data team at an online publishing company. One of your verticals is a food publication. A product manager on your team wants to build a feature for this vertical that enables users to query by cuisine, not just by ingredients. Most of your recipes are unlabeled, and it's infeasible to label them by hand. Luckily, you have a small training set of about 10,000 recipes with labeled cuisines.

* Design and execute a method to predict the cuisine of a recipe given only its ingredients.
  + Data Due Diligence: All-Purpose Flour and Flour are likely the same ingredient, but red onions and yellow onions are incredibly different.
* For each major cuisine, what are the driving ingredients that characterize it? What are the features of a cuisine that drive misclassification in your method above?
* How could you design this to be robust enough to understand similarities / substitutions between ingredients?
* Your product manager indicates a likelihood that you will only need to write a guideline for an outsourced team to hand label the remaining corpus. How would you go about writing this guide for a few major cuisines?