Whats Cooking: Capstone Project

Mallesh

January 5, 2016

Introduction

This project has been picked from one of the Kaggle competitions. Even though this might not address a real problem (or may be it does?), it has a lot of scope for learning and similar principles can be applied to several other classification problems. The primary objective of the project is to predict the type of cuisine based on the list of ingredients used in the dish. A classification model will be developed based on a training dataset and then be used to predict the cuisines in a test dataset.

Datasets can be found here:

https://www.kaggle.com/c/whats-cooking/data

Download both train.json and test.json into the working directory.

```r
require(jsonlite)

cookTrain <- fromJSON("train.json", flatten = TRUE)
cookTest <- fromJSON("test.json", flatten = TRUE)
```

Lets look at the structure of the data.

```r
str(cookTrain)
```

```
## 'data.frame':    39774 obs. of  3 variables:
##  $ id         : int  10259 25693 20130 22213 13162 6602 42779 3735 16903 12734 ...
##  $ cuisine    : chr  "greek" "southern_us" "filipino" "indian" ...
##  $ ingredients:List of 39774
##   ..$ : chr  "romaine lettuce" "black olives" "grape tomatoes" "garlic" ...
##   ..$ : chr  "plain flour" "ground pepper" "salt" "tomatoes" ...
##   ..$ : chr  "eggs" "pepper" "salt" "mayonaise" ...
##   ..$ : chr  "water" "vegetable oil" "wheat" "salt"
##   ..$ : chr  "black pepper" "shallots" "cornflour" "cayenne pepper" ...
##   ..$ : chr  "plain flour" "sugar" "butter" "eggs" ...
##   ..$ : chr  "olive oil" "salt" "medium shrimp" "pepper" ...
##   ..$ : chr  "sugar" "pistachio nuts" "white almond bark" "flour" ...
##   ..$ : chr  "olive oil" "purple onion" "fresh pineapple" "pork" ...
##   ..$ : chr  "chopped tomatoes" "fresh basil" "garlic" "extra-virgin olive oil" ...
##   ..$ : chr  "pimentos" "sweet pepper" "dried oregano" "olive oil" ...
##   ..$ : chr  "low sodium soy sauce" "fresh ginger" "dry mustard" "green beans" ...
##   ..$ : chr  "Italian parsley leaves" "walnuts" "hot red pepper flakes" "extra-virgin olive oil" ...
##   ..$ : chr  "ground cinnamon" "fresh cilantro" "chili powder" "ground coriander" ...
##   ..$ : chr  "fresh parmesan cheese" "butter" "all-purpose flour" "fat free less sodium chicken broth
##   ..$ : chr  "tumeric" "vegetable stock" "tomatoes" "garam masala" ...
##   ..$ : chr  "greek yogurt" "lemon curd" "confectioners sugar" "raspberries"
##   ..$ : chr  "italian seasoning" "broiler-fryer chicken" "mayonaise" "zesty italian dressing"
##   ..$ : chr  "sugar" "hot chili" "asian fish sauce" "lime juice"
##   ..$ : chr  "soy sauce" "vegetable oil" "red bell pepper" "chicken broth" ...
##   ..$ : chr  "pork loin" "roasted peanuts" "chopped cilantro fresh" "hoisin sauce" ...
##   ..$ : chr  "roma tomatoes" "kosher salt" "purple onion" "jalapeno chilies" ...
##   ..$ : chr  "low-fat mayonnaise" "pepper" "salt" "baking potatoes" ...
##   ..$ : chr  "sesame seeds" "red pepper" "yellow peppers" "water" ...
##   ..$ : chr  "marinara sauce" "flat leaf parsley" "olive oil" "linguine" ...
##   ..$ : chr  "sugar" "lo mein noodles" "salt" "chicken broth" ...
##   ..$ : chr  "herbs" "lemon juice" "fresh tomatoes" "paprika" ...
##   ..$ : chr  "ground black pepper" "butter" "sliced mushrooms" "sherry" ...
##   ..$ : chr  "green bell pepper" "egg roll wrappers" "sweet and sour sauce" "corn starch" ...
##   ..$ : chr  "flour tortillas" "cheese" "breakfast sausages" "large eggs"
##   ..$ : chr  "yellow corn meal" "boiling water" "butter" "fresh parmesan cheese" ...
##   ..$ : chr  "chicken broth" "chicken breasts" "hot sauce" "red bell pepper" ...
##   ..$ : chr  "chili powder" "crushed red pepper flakes" "garlic powder" "sea salt" ...
##   ..$ : chr  "eggs" "shallots" "firm tofu" "beansprouts" ...
##   ..$ : chr  "olive oil" "onions" "crushed garlic" "dried oregano" ...
##   ..$ : chr  "olive oil" "diced tomatoes" "Johnsonville Andouille Dinner Sausage" "parsley" ...
##   ..$ : chr  "olive oil" "bread slices" "great northern beans" "garlic cloves" ...
##   ..$ : chr  "chicken broth" "cooking oil" "chinese five-spice powder" "ground black pepper" ...
##   ..$ : chr  "green onions" "cream cheese" "shredded cheddar cheese" "cayenne pepper" ...
##   ..$ : chr  "collard greens" "extra-virgin olive oil" "ham hock" "chicken stock" ...
##   ..$ : chr  "Oscar Mayer Deli Fresh Smoked Ham" "hoagie rolls" "salami" "giardiniera" ...
##   ..$ : chr  "ice cubes" "club soda" "white rum" "lime" ...
##   ..$ : chr  "cooked chicken" "enchilada sauce" "sliced green onions" "picante sauce" ...
##   ..$ : chr  "salmon fillets" "shallots" "cumin seed" "fresh cilantro" ...
##   ..$ : chr  "tomatoes" "chicken breast halves" "chopped cilantro fresh" "white vinegar" ...
##   ..$ : chr  "eggs" "mandarin oranges" "water" "orange liqueur" ...
##   ..$ : chr  "sugar" "salt" "fennel bulb" "water" ...
##   ..$ : chr  "sugar" "all-purpose flour" "vegetable oil" "white cornmeal" ...
##   ..$ : chr  "butter" "crab boil" "garlic" "old bay seasoning" ...
##   ..$ : chr  "mayonaise" "white sugar" "ground black pepper" "salt" ...
##   ..$ : chr  "sirloin" "mirin" "yellow onion" "low sodium soy sauce" ...
##   ..$ : chr  "large eggs" "whipping cream" "chicken broth" "ground red pepper" ...
##   ..$ : chr  "fresh basil" "bay leaves" "crushed red pepper" "mussels" ...
##   ..$ : chr  "sugar" "large eggs" "all-purpose flour" "baking soda" ...
##   ..$ : chr  "lemon" "pesto" "salmon fillets" "white wine"
##   ..$ : chr  "bread crumbs" "unsalted butter" "onion powder" "curry" ...
##   ..$ : chr  "melted butter" "matcha green tea powder" "white sugar" "milk" ...
##   ..$ : chr  "coarse salt" "fenugreek" "urad dal" "potatoes" ...
##   ..$ : chr  "pizza crust" "plum tomatoes" "pesto" "part-skim mozzarella cheese"
##   ..$ : chr  "cooking spray" "salt" "black pepper" "yukon gold potatoes" ...
##   ..$ : chr  "sugar" "chicken thighs" "cooking oil" "fish sauce" ...
##   ..$ : chr  "lemongrass" "large garlic cloves" "rice" "unsweetened coconut milk" ...
##   ..$ : chr  "chicken legs" "chile pepper" "ghee" "tomato paste" ...
##   ..$ : chr  "stock" "curry powder" "cracked black pepper" "minced beef" ...
##   ..$ : chr  "crushed tomatoes" "garlic" "fresh rosemary" "ground black pepper" ...
##   ..$ : chr  "extra firm tofu" "coconut milk" "fresh basil" "red curry paste" ...
##   ..$ : chr  "jasmine rice" "garlic" "scallions" "sugar" ...
##   ..$ : chr  "vanilla" "milk" "large egg yolks" "sugar" ...
##   ..$ : chr  "orange juice concentrate" "pumpkin purée" "marshmallow creme" "toasted pecans" ...
##   ..$ : chr  "black pepper" "white sugar" "white vinegar" "salt" ...
##   ..$ : chr  "large eggs" "serrano ham" "manchego cheese" "butter" ...
##   ..$ : chr  "burger buns" "fresh cilantro" "chili powder" "garlic cloves" ...
##   ..$ : chr  "ground cloves" "whole nutmegs" "ground ginger" "ground coriander" ...
##   ..$ : chr  "tomatoes" "red pepper" "olive oil" "Italian bread" ...
##   ..$ : chr  "sausage casings" "ground black pepper" "garlic" "honey" ...
##   ..$ : chr  "tumeric" "olive oil" "lemon" "saffron" ...
##   ..$ : chr  "ground pepper" "paprika" "ground cardamom" "chopped cilantro fresh" ...
##   ..$ : chr  "sweetened condensed milk" "ice" "espresso"
##   ..$ : chr  "top round steak" "vegetable oil" "shiitake" "soy sauce" ...
##   ..$ : chr  "avocado" "chopped cilantro fresh" "jalapeno chilies" "finely chopped onion" ...
##   ..$ : chr  "pecans" "golden brown sugar" "crumbled blue cheese" "garlic cloves" ...
##   ..$ : chr  "sugar" "vanilla" "eggs" "self rising flour" ...
##   ..$ : chr  "Madeira" "foie gras" "demi-glace" "sherry vinegar" ...
##   ..$ : chr  "fenugreek leaves" "olive oil" "garlic" "black mustard seeds" ...
##   ..$ : chr  "black peppercorns" "cinnamon sticks" "cardamom pods" "cumin seed" ...
##   ..$ : chr  "spinach" "asiago" "whole wheat pasta" "olive oil" ...
##   ..$ : chr  "chestnuts" "granulated sugar" "whole milk ricotta cheese" "coffee ice cream" ...
##   ..$ : chr  "baby spinach leaves" "naan" "unsalted butter" "chopped garlic" ...
##   ..$ : chr  "water" "barley"
##   ..$ : chr  "cooked ham" "red bell pepper" "seasoning" "potatoes" ...
##   ..$ : chr  "salt" "starchy potatoes" "grated nutmeg" "flour"
##   ..$ : chr  "ground black pepper" "all-purpose flour" "milk" "buttermilk" ...
##   ..$ : chr  "granulated sugar" "all-purpose flour" "vegetable shortening" "baking powder" ...
##   ..$ : chr  "ground pork" "finely chopped fresh parsley" "onions" "salt" ...
##   ..$ : chr  "ground cinnamon" "whole milk" "golden brown sugar" "heavy whipping cream" ...
##   ..$ : chr  "boneless chicken skinless thigh" "lime" "epazote" "bay leaf" ...
##   ..$ : chr  "catfish fillets" "ground black pepper" "salt" "dried thyme" ...
##   ..$ : chr  "olive oil" "sea salt" "coconut milk" "water" ...
##   ..$ : chr  "olive oil" "olives" "salt" "blood orange" ...
##   .. [list output truncated]
```

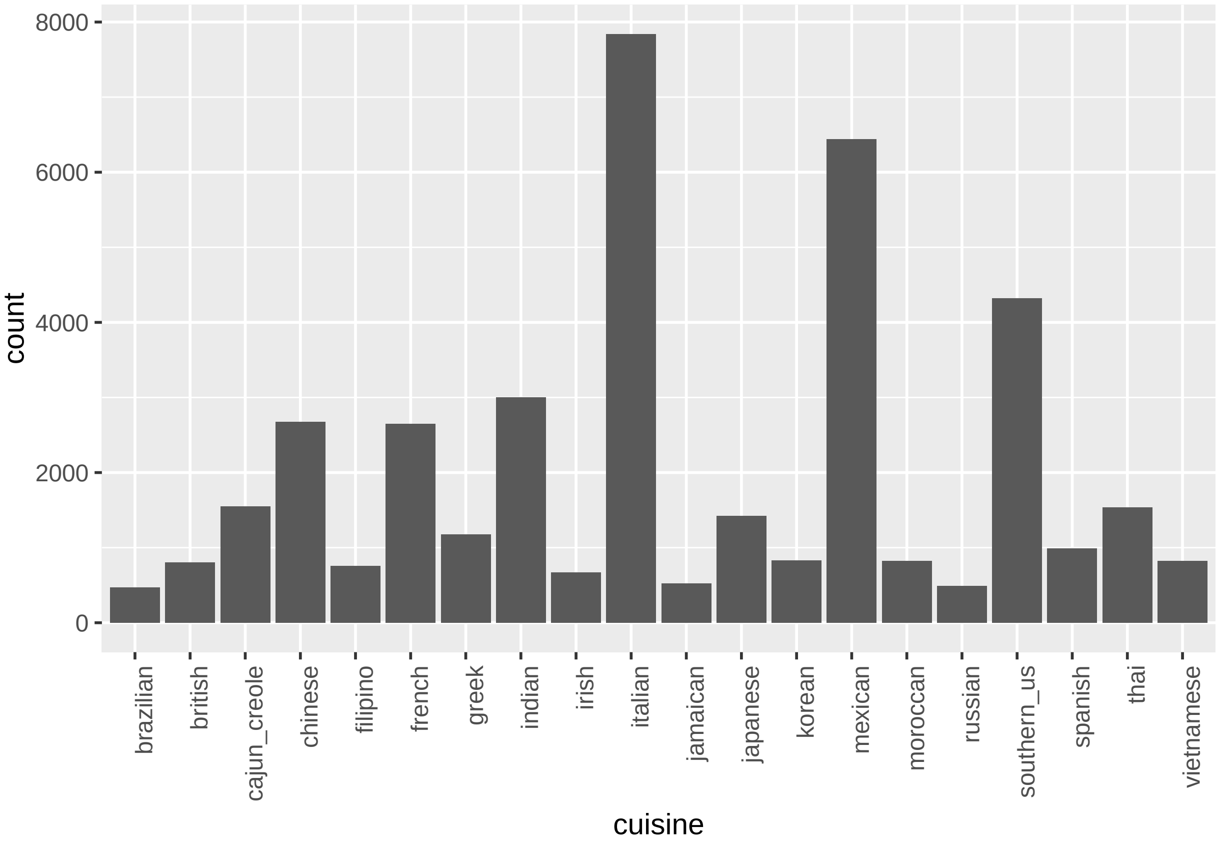
Looking at the structure of the file, we know that there are three attributes, id - indicates a dish ID (from kaggle), cusine - indicates a type of cuisine and ingredients - a list of ingredients.

Also, looking at some of the values displayed here, it can be observed that there are several variations of same kind of ingredients. Which indiactes there is some level of cleaning need to be done to be able build a good model.

Let us look at the number of recipes in each cusine to see how the data is spread:

```r
require(ggplot2)

ggplot(data = cookTrain, aes(x = cuisine)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



The graph indicates that there a lot of recipes in Italian. So, lets mark all the recipes in test file as Italian as a benchmark. Also, add indicators to both test and training datasets, so that we can combine and split them as needed.

```
cookTrain$type <- "train"

cookTest$type <- "test"

cookTest$cuisine <- "italian"
```

Create a combined dataset called cookCombined, this would enable us to build the sparse matrix without the problem of having new variables from test dataset. These datasets would be spearated before building the model.

```
cookCombined <- rbind(cookTrain, cookTest)
```

Data Wrangling and clean up:

It is observed that some of the ingredients have measurements like 1 oz etc and some of the ingredients have punctuations like . and parentheses.

create a corpus of ingredients for further processing. Using TM package:

```
require(tm)
require(wordcloud)

bagOfIng <- Corpus(VectorSource(cookCombined$ingredients))
```

Now remove the punctuations and stop words using the tm_map function.

```
bagOfIng <- tm_map(bagOfIng, stemDocument)
bagOfIng <- tm_map(bagOfIng, removePunctuation)
bagOfIng <- tm_map(bagOfIng, content_transformer(tolower))
bagOfIng <- tm_map(bagOfIng, removeNumbers)
bagOfIng <- tm_map(bagOfIng, function(x) removeWords(x, c("the","frozen","fresh","oz")))
```

Stem the document terms and create a term matrix:

```
bagOfIngDTM <- DocumentTermMatrix(bagOfIng)
```

We can plot some word clouds to see what are some of the common terms across cuisines and some of the least appearing terms. This way we can also find any anomolies.

Least Common terms:

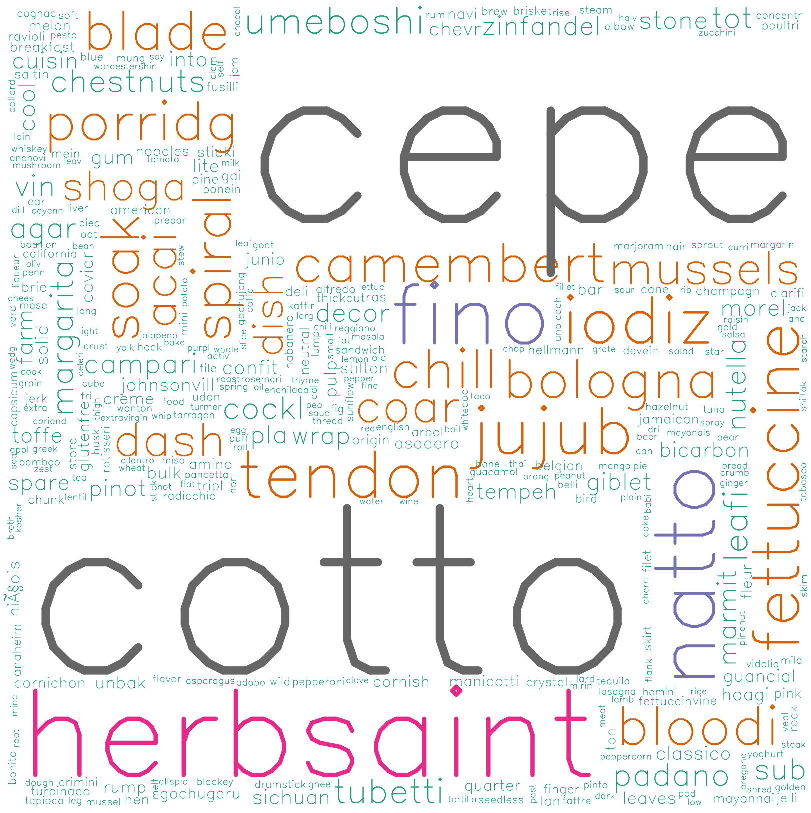
```
matrix <- as.data.frame(as.matrix(bagOfIngDTM))

aggregation <- sort(colSums(matrix),decreasing=TRUE)

df <- data.frame(word = names(aggregation),freq=1/aggregation)

pal <- brewer.pal(8, "Dark2")


wordcloud(df$word,df$freq, scale=c(8,.3),min.freq=2,max.words=Inf, random.order=T, rot.per=.15, colors=
```

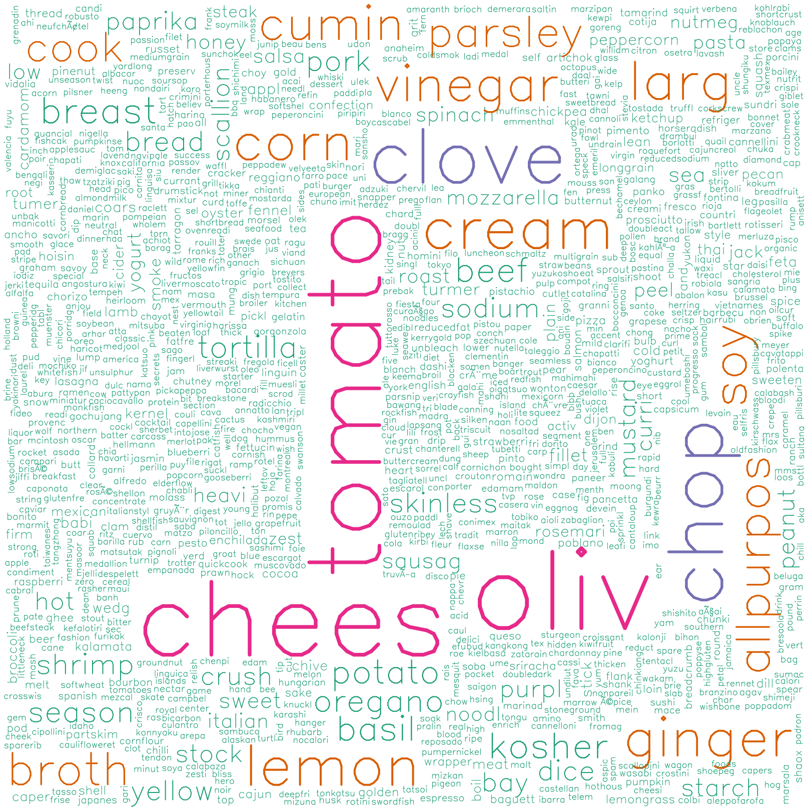


Most common terms:

```r
df <- data.frame(word = names(aggregation),freq = aggregation)

pal <- brewer.pal(8, "Dark2")


wordcloud(df$word,df$freq, scale=c(8,.3),min.freq=2,max.words=Inf, random.order=T, rot.per=.15, colors=
```



To Do:

Based on the word clouds above, we can clean up some ingredients, like removing the company names like “Foster” etc.

Convert the matrix into a data frame and split the training and testing datasets.

```r
bagOfIngDTM <- as.data.frame(as.matrix(bagOfIngDTM))

bagOfIngDTM$type <- as.factor(cookCombined$type)

#inTrain <- createDataPartition(y = ingredientsDTM$cuisine, p = 0.6, list = FALSE)
training <- bagOfIngDTM[bagOfIngDTM$type == "train",]
testing <- bagOfIngDTM[bagOfIngDTM$type == "test",]

training$cuisine <- as.factor(cookTrain$cuisine)
testing$cuisine <- as.factor("italian")
testing$id <- cookTest$id
```

Build the model using rpart package:

```
require(rpart)

model <- rpart(cuisine ~ ., data = training, method = "class")
```

Applying the model on Test data:

```
pred <- predict(model, newdata = testing, type = "class")

testing$cuisine <- pred
```

Write the predicted values into a CSV file for submission on kaggle:

```
write.csv(testing[,c("id","cuisine")], "outputCooking.csv", quote=FALSE,
          row.names = FALSE)
```

Results indicate a prediction accurcy at 0.40.

Few other things being pursued to improve the Model performance:

Use Random Forest for model building Apply Gradient Boost Clean up the ingredients to remove some unnecessary terms like company names etc.