Open Food Database: Ingredient Co-Occurence Network

Kate Cooper

Data Download

Data was downloaded from https://world.openfoodfacts.org/data (https://world.openfoodfacts.org/data) on 03-21-2019 at 9:57am as a CSV file.

```
## Set variables
workingDir = "/Users/katedempsey/Documents/Research/UNO/CooperLab/ingredient_network/"

## Change to the data folder of your working directory and download
setwd(paste0(workingDir,"data"))
ret = download.file("https://static.openfoodfacts.org/data/en.openfoodfacts.org.product
s.csv","raw_data.csv")
```

Environment setup:
1. Set your working directory as the variable workingDir
2. The code assumes in your working directory, you have subfolders: *workingDir/data* and *workingDir/results*

At time of download (Mar 2019), the CSV file downloaded was 2.21GB, so plan accordingly. The file itself contains 798,919 lines by wordcount and is tab-delimited.

```
workingDir = "/Users/katedempsey/Documents/Research/UNO/CooperLab/ingredient_network/"
setwd(paste0(workingDir,"data"))
system('wc -l raw_data.csv')
```

Pre-processing

The dataset itself contains a lot of information; at this exploratory stage we only would like to investigate ingredients per product. We also have to check the following: * Are there dupliate products in the data? * Are all ingredients named in the same way?

First, we will check if there are duplicates by barcode, the first column in the dataset. We will search only for foods sold in the United States (column 32). We also want to include the ingredients (columnt 35), and any allergens that may be noted (columns 36 and 37).

```
system('cut -f 1,32,35,36,37 raw_data.csv | grep \'United States\' | uniq | wc -l')
```

We want to look at product by barcode as a unique ID for the product and we will make our ingredient network by comparing ingredients from the “ingredients” text in column 35, so in the next steps we extract columns 1 and 35 only from the data, and remove duplicates. This cuts our file down to a much more manageable size of 38.5MB.

```
system('cut -f 1,32,35,36,37 raw_data.csv | grep \'United States\' | uniq > raw_ingredie
nts.txt')
```

Interestingly, we note that there are 175,163 rows (unique barcodes for foods in the United States) listed in our raw_ingredients file, but only 1,419 of them have allergens specified in the database. This represents 0.81% of foods listed in our unique United States dataset. It is assumed that entries into the Open Food Database are not reviewed for correctness [citation needed], but the 2004 Food Allergen Labeling Consumer Protection Act (FALCPA), which took effect in January 2006, requires all food labels in the United States to identify if a product contains one of the eight major allergens. (Source (https://www.fda.gov/food/guidanceregulation/guidancedocumentsregulatoryinformation/allergens/ucm106890.htm))

```
system('wc -l raw_ingredients.txt')
system('cut -f 4 raw_ingredients.txt  | uniq | wc -l')
```

Next, we want to remove any barcodes that do not have ingredients associated with them. To do this, lets read the file into R and begin manipulating it in memory.

```
raw_ingredients = read.csv("raw_ingredients.txt",header = TRUE)
```

This should result in a dataframe called *raw_ingredients* that contains 499,879 observations of 2 variables. Next, we remove duplicates by removing rows for which *ingredients_text* is empty.

```
ingredients <- raw_ingredients[-which(raw_ingredients$ingredients_text == ""), ]

# You can now remove the raw_ingredients variable if you are feeling confident in the re
producibility of your project
remove(raw_ingredients)
```

Once this is complete, we can begin to investigate and compare ingredients. Taking a brief look at the data, the first challenge to overcome is the case of the text; evaluating “Salt” and “salt” as equal ingredients will be easier if they are written in the same case. So next we change the ingredients list to all lowercase.

```
ingredients$ingredients_text <- tolower(ingredients$ingredients_text)
```

Certain ingredients are followed by a percentage (i.e. “milk chocolate 32.7%”). For purposes of building our co-occurence network, we will disregard these percentages and remove them, both for US (32.7%) and European (32,7%) formatting styles.

```
ingredients$ingredients_text <-gsub('\\d+[.,]*\\d*\\s{0,1}%','',ingredients$ingredients_
text)
```

Further, we want to make similar changes to make comparability of ingredients similar.

```
# Remove the term "organic" from ingredients as organic is not regulated by the FDA
ingredients$ingredients_text <-gsub('organic ','',ingredients$ingredients_text)

# Remove special letters and characters and replace with standard [a-z] or otherwise
ingredients$ingredients_text <-gsub('[éèë]','e',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('ï','i',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('[âà]','a',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('ô','o',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('_','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('?','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('\\[\\]','',ingredients$ingredients_text)


# Remove the term "ingredients" as it is redundant
ingredients$ingredients_text <-gsub('ingredient[s]*\\s{0,1}\\:*','',ingredients$ingredie
nts_text)
ingredients$ingredients_text <-gsub('ingrédient[s]*\\s{0,1}\\:*','',ingredients$ingredie
nts_text)
ingredients$ingredients_text <-gsub('ingrédiente[ns]*\\s{0,1}\\:*','',ingredients$ingred
ients_text)
ingredients$ingredients_text <-gsub('amount','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('serving','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('nourishment','',ingredients$ingredients_text)

# Remove preparatory or provenance terms that do not affect molecular makeup of the ingr
edient (but may remove bacteria/pathogens)
ingredients$ingredients_text <-gsub('cultured','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('pasteurized','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('distilled','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('california','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('grade a+','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('extra ','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('virgin ','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('free range','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('french','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('whole ','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('rolled ','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('expeller ','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('pressed ','',ingredients$ingredients_text)
ingredients$ingredients_text <-gsub('cow\'s milk','milk',ingredients$ingredients_text)
```

Finally, we want to take a look at ingredients which are comma-separated. For each unique barcode, we will name each ingredient as a node in our network and will draw an edge between them, indicating that they co-occur in that food item, resulting in a K_n network for each food item where *n* = the number of ingredients for that food item.

We will later collectively look at the network as a whole, where

Exploratory Network Analysis

- What are the most commonly co-occuring ingredients used in the Open Food Facts database?
- What are the structures of commonly co-occuring ingredients used in the Open Food Facts database?

- Are the top 8 allergens commonly directly co-occurring ($k <= 1$) or not?

Targets

The Journal of AOAC International is dedicated to publishing basic and applied research in the analytical sciences related to foods, drugs, agriculture, and the environment. Emphasis is on research and development of precise, accurate, sensitive methods of analysis in the following areas: ‘Animal and Plant Nutrition, Health, and Safety’ Dietary Supplements, and Food Chemical Contaminants ‘Drug Formulations and Clinical Methods’ Food Biological Contaminants, and Microbiological Methods ‘Food Composition and Additives, and Infant Formula and Adult Nutritionals’ Residues and Trace Elements ‘Statistical Analysis and Chemometrics’ Veterinary Drug Residues The Journal is the forum for the exchange of information among method researchers who must keep informed of new technology and techniques in industry. The Journal publishes the fully refereed reports on developing, improving, and testing uniform, precise, and accurate methods.

The Journal is unique in that it also publishes types of papers not found in any other analytical science journal: Standard Method Performance Requirements (SMPRs’), AOAC First and Final Action Official Methods, and stakeholder panel related single- and multi-laboratory validation reports, as well as the validation reports from AOAC’s Performance Tested Method program, providing independent review of test kit performance. The publication of ‘stakeholder output,’ an essential function of the Journal, is scheduled throughout the year.

For manuscript guidelines, visit: http://www.aoac.org/aoac_prod_imis/aoac_docs/journal/author_resources.pdf (http://www.aoac.org/aoac_prod_imis/aoac_docs/journal/author_resources.pdf)

OR

http://u7.ift.org/knowledge-center/read-ift-publications/journal-of-food-science.aspx (http://u7.ift.org/knowledge-center/read-ift-publications/journal-of-food-science.aspx)

Or

Innovative Food Science & Emerging Technologies

Or

Molecular Nutrition and Food Research

Or

Food Research International