What's Cooking? Classification of recipes into cuisines according to ingredients

Source of data

Kaggle (unknown author), with dataset compiled from Yummly

Description of the dataset

The dataset was downloaded from the American recipe website Yummly (www.yummly.com) and contains almost 40,000 recipes and 6700 ingredients. Each recipe is classified (by their author) into one of the cuisine categories (e.g. British, Japanese, Mexican), and each contains a set of ingredients needed to cook given food (e.g. water, vegetable oil, wheat, salt; no quantities are given). I downloaded the original JSON format file from Kaggle and prepared three versions of tables for further use, both having columns representing individual ingredients. In the first, complete version (with > 40,000 rows), each row represents the original recipe, with a value of 1 for each ingredient present in that recipe. In the second, aggregated table version (with 20 rows), each row contains one cuisine, and values in the cells represent counts of how many times in recipes from a given cuisine a given ingredient was used. In the third, reduced table version, I first converted the counts of recipes into relative numbers per cuisine (by dividing each element of the second version table by row sum), and selected only ingredients that occur in at least 1% of all recipes (column sum is larger than 0.01).

Locality

Global (I assume), but perhaps with geographic bias (since www.yummly.com is an American server)

Data for download

File name	File type	Description
recipes_long.txt	tab-delimited txt format	All recipes in a long format, containing recipe ID, cuisine, and ingredients. Need to be reshaped into wide format for analysis (code included in the section for direct upload of data into R)
recipes_cuisines.txt	tab-delimited txt format	Contains Cuisines × ingredients table (20 cuisines in rows, 6714 ingredients in columns, cell values equal to the count of recipes of give cuisine mentioning given ingredient)
recipes_cuisines_select.txt	tab-delimited txt format	Contains reduced cuisines × ingredients table (20 cuisines in rows, 1693 ingredients in columns, cell values equal to the relative proportion of recipes of give cuisine mentioning given ingredient); only ingredient occurring in at least 1% of all recipes are included. Columns with ingredients are ordered from the most common to the least common (decreasing column sums).
recipes_cuisines_count.txt	tab-delimited txt format	Contains two column, cuisine and the count of recipes in the dataset assigned to this cuisine.

Script for direct import of data to R

Cuisine x ingredients table and recipe count table

```
recipes_cuisines <- read.delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/recipes_cuisin
es.txt')
recipes_cuisines_select <- read.delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/recipes_cuisin
es_select.txt', row.names = 1)
recipes_cuisines_count <- read.delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/recipes_cuisin
es_count.txt')
```

Individual recipes (long format) and reshaping into wide format

```
library (readr)
recipes_long <- read_delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/recipes_long.t
xt')

recipes_compl_tb <- as_tibble (recipes_long) %>%
  select (-cuisine) %>%
  mutate (presence = 1L) %>%
  pivot_wider (names_from = ingredients, values_from = presence, values_fn = sum,
values_fill = 0L)

recipes_compl <- as.data.frame (recipes_compl_tb)
rownames (recipes_compl) <- recipes_compl_tb$id
recipes_compl <- recipes_compl [,-1]
recipes_compl_cuisine <- distinct (recipes_long[,-3])
```

The code used for reshaping the long data from `recipes_long` results in two objects: `recipes_compl`, a data.frame containing each recipe as a row and each ingredient as a column, with cell values a presence-absence data; and `recipes_compl_cuisine` (a tibble) with recipe ID and cuisine to which the recipe was assigned on yummly.com.

References

- Dataset from Kaggle, downloaded from Yummly (yummly.com). Unknown author.

From:
https://anadat-r.davidzeleny.net/ - **Analysis of community ecology data in R**

Permanent link:
https://anadat-r.davidzeleny.net/doku.php/en:data:recipes

Last update: **2023/05/07 09:22**