Name of your project

Redouane Elghazi redouane.elghazi@epfl.ch
Pierre Mahmoud–Lamy pierre.mahmoud-lamy@epfl.ch
Enguerrand Prebet enguerrand.prebet@epfl.ch

Abstract

Can recipes be grouped into meaningful classes in terms of ingredients? Are these groups characterized by their distributions in carbs,fat and proteins?

1 Database Presentation

We gather the data from (Robert West et al., 2013). This data contains recipes from several cooking websites. A recipe entry have informations such as the calories (fat, proteins, carb) and the ingredients.

To only extract the ingredients from the list, we used an external data base from http://openfood.schoolofdata.ch/food-composition-ch/ as a starting point. It contains foods but also a food group such as "vegetables" or "cereals".

We had to append words that were not in the database but that appears in recipes, especially cheese names, pasta names and different types of sauce.

2 Processing the database

To only extract foods ingredients and not quantities or others words, we consider our external database with words sorted by decreasing length. For each ingredient, we then process through the recipes and check if it appears in the string of ingredients. In that case, we add it to our list of ingredients and delete it from the string.

By sorting and deleting ingredients, we ensure that combination of words like "black pepper" is found without also finding "pepper". Some words were still matched incorrectly, with "tea" found in recipes because of the use of "teaspoon", and "gin" found in "virgin" for instance. So before looking for our ingredients, we preprocess the recipe's ingredients by removing those words first. We also took into consideration plural and separators.

When looking at the nutritional informations, we saw that some recipes are too energetic to be normalized per serving, being over the recommended amount for one person **per day**. To be able to compare different recipes, we normalized our data by the total energy for the recipe, and thus considered proportion of fat, carb and proteins instead of their raw value.

3 Clustering

To reorganize our different recipes into groups according to their ingredients, we need a decent representation. Our choice was to consider boolean value for each food group if a recipe contains an ingredients from that food group. This gives us a N-dimensional vector with value in $\{0,1\}^N$, with N being the number of food group (11 in our case).

We use a K-means clustering algorithm to create our recipe groups. Starting from K random means, we assign each recipe to the closest mean (the distance being defined as the sum of differences over the N elements). Then each mean is updated to correspond to the effective mean of the elements assigned to this mean.

In practice, though recipes get values in $\{0,1\}^N$, the means have real values in $[0,1]^N$ to get the best possible clustering.

Moreover, after having tested different values of K to study the convergence of the groups, we set the option to fix the initial means, for reproducibility.

4 First results

After several tests, the more meaningful number of groups seems to be 4. By meaningful, we mean that we can visualize what the groups represent. Recall that the output of the algorithm is the assignement vector for each recipe, and the mean

vectors. We then need to interpret the meaning of those means:

- the first group represents desserts based on complex preparations, mostly with eggs and flour or other cereal base, such as pies, cakes, muffins,...
- the second group represents more simple desserts, with fewer ingredients and more fruit-based, such as fruit salads, baked apples,... but also chocolate mousse,...
- the third group consists in complex prepared meals, with meat and vegetables. It uses a lot of different food groups (eggs, dairy, fats) and can for example be stews, boeuf bourguignon,...
- the last group is easy-to-do dishes with meat and vegetables. It uses more condiments (mustard, sauces,...) and can be bolognese pasta, hot-dogs,...

The split of the groups can be considered using 2 questions:

- Is this a main course ?
- Is it simple ?

5 Classification by the nutriments

- analysis of distributions - propensity score

6 Conclusions

References

Robert West, Ryen W. White, and Eric Horvitz. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proceedings of the 22nd International World Wide Web Conference*, 2013.