Smart recipes

Omar Mehio
omar.mehio@epfl.ch

Leonardo Perrone
leonardo.perrone@epfl.ch

Attila Bekker
attila.bekker@epfl.ch

Abstract

The aim of this document is to find correlations between attributes and characteristics of food recipes and the corresponding ratings and reviews given to that recipe.

Recipes are gathered from the most popular online recipe websites, extracting common attributes and normalizing them between the different representations. Apart from the attributes explicitly included in the recipe descriptions, new properties are being defined by grouping recipe ingredients into different categories. Furthermore, sentiment analysis is performed on the content of recipe reviews, providing a way to determine the corresponding ratings.

Performing a range of different analysis techniques provides a set of hints and best practices to help recipe authors to maximize their ratings by choosing the recipes with attributes most likely to be rated positively.

1 Introduction

This document has been prepared for the **Applied Data Analytics** course at EPFL in the Fall of 2018.

2 Data acquisition

The recipe data has been extracted from the *Cooking recipes* database, which is a collection of HTML documents from a wide variety of websites.

2.1 Data sources

The analysis has been based on data extracted from the most popular websites containing food recipes. The data included recipes from over 50 websites, all featuring their own custom format and attributes. Since each website required a custom scraper to be built, and each website displays different attributes of recipes and their ingredients, the aim was to select a small number of websites which still offer a good representation of the entire data set. The aim was to work with 2⁄3 of the total number of available recipes.

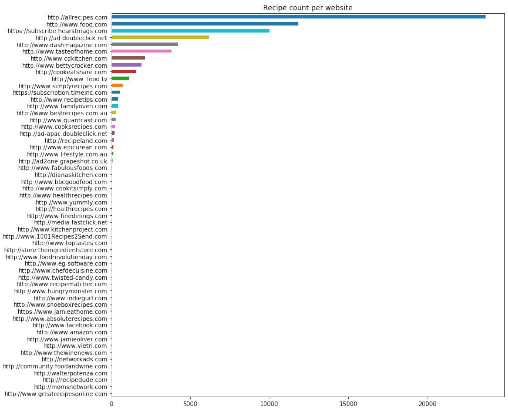

Figure 1. Recipe count

The top 3 websites containing the most recipes cover about 65% of all data:

1. http://allrecipes.com
2. http://www.food.com
3. https://subscribe.hearstmags.com (food-net)

2.2 Data import

The three selected websites provided different recipe attributes in different formats, which needed to be parsed and consolidated into a common format in order to perform meaningful analysis on them. Common attributes were:

- name
- review text

- review date
- rating
- category
- ingredient list
- nutritional values
- preparation time
- cooking time
- difficulty
- serving size

These attributes may have different scales, different measurement units, or may be missing altogether.

The three websites required custom scraping algorithms. The source data volume was about 13 GB, making it feasible to be processed on a single computer. The scraping took 20-30 minutes, the results were stored in *json* format to allow efficient processing.

2.3 Data cleaning

Due to the different formats, the data sets had to be cleaned separately. Nutrition information had to be converted to numeric values on a unified scale, using the same measurement units across all recipes. Time values expressed in hours or days needed to be converted to minutes. Rating values had to be scaled to a common range. Some text values had to be converted to numeric values.

2.4 Review pre-processing

The content of the reviews had special importance for the analysis. In order to analyze the text, a pipeline was constructed to clear the text. Numbers and other non-alphanumeric characters were removed, and all remaining text was converted to lowercase. Stop words offered little value in terms of semantic analysis, so they were removed. Lemmatizing removed prefixes and suffixes, helping to group similar words. Finally a document term matrix was constructed using both TF-IDF and one-hot encoding matrices.

3 Preliminary visual analysis

Once the raw data had been converted into a clean data set with common attributes, they were visualized in order to better understand the characteristics. This analysis was performed before combining the data sets into a single one, allowing comparison between websites.

3.1 Ratings

Ratings were found to be dominated by high values. After converting them to a common scale of 3 it was observed that good ratings highly outnumber bad or medium ones.

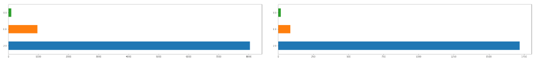

Figure 2. Ratings distribution

3.2 Serving size

Serving sizes ranged up to 300, but there were only a small number of recipes for more than 50 persons. The 75% percentile fell to serving sizes of 8 and 12 respectively for the two websites with the most recipes.

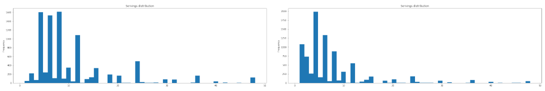

Figure 3. Servings distribution

3.3 Reviews over time

Most reviews have been created between 2000 and 2010, indicating the period when the data had been collected. The density curves suggested different trends for the popularity of the websites. The allrecipes site showed declining review numbers over the years, while food.com had fairly constant activity, and food-net reviews sharply increased towards the end of the period.

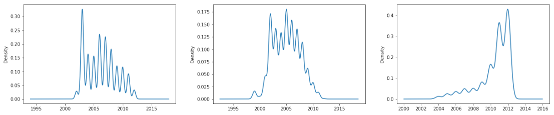

Figure 4. Reviews over time

3.4 Cooking time

Cooking times ranged from a few minutes to several hours, but the distribution showed that most recipes required less than 1 hour cooking, recipes requiring 30 minutes or less being the most common.

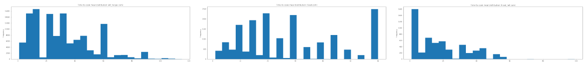

Figure 5. Cooking time distribution

4 Ingredient clustering

Comparing recipes based on ingredients is problematic, due to the high number of possible ingredients. By grouping similar ingredients into clusters we were able classify recipes by how many ingredients they contained from each cluster. The following procedure has been performed separately for the different websites, for the sake of simplicity we demonstrate the results from food.com only.

First, measurements and other non-food related words needed to be removed from the review text. For example, words like cup or tablespoon were featured frequently, but were irrelevant to define ingredient clusters. Removing these words and keeping only ingredients resulted in more balanced distribution.

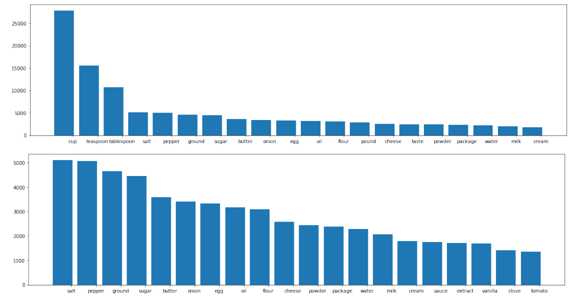

Figure 6. Ingredient distribution

By applying pre-trained word2vec vectors, we mapped each ingredient into a word vector, and projected them to a 2-dimensional plane using t-distributed stochastic neighbor embedding.

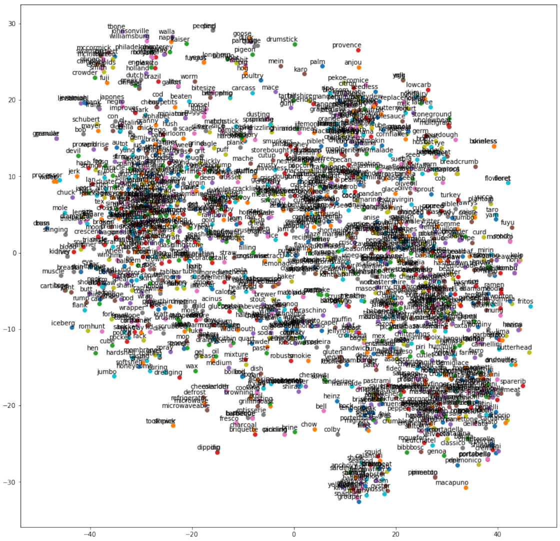

Figure 7. Ingredient word vectors

We used silhouette analysis to find the optimal number of clusters for the k-means clustering, determining that 29 clusters produced the best performance.

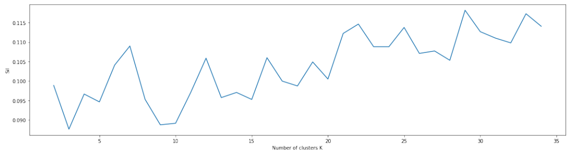

Figure 8. K-means silhouette analysis

Based on manual analysis of the classified ingredients, the individual clusters have been named accordingly and projected on the 2D plane.

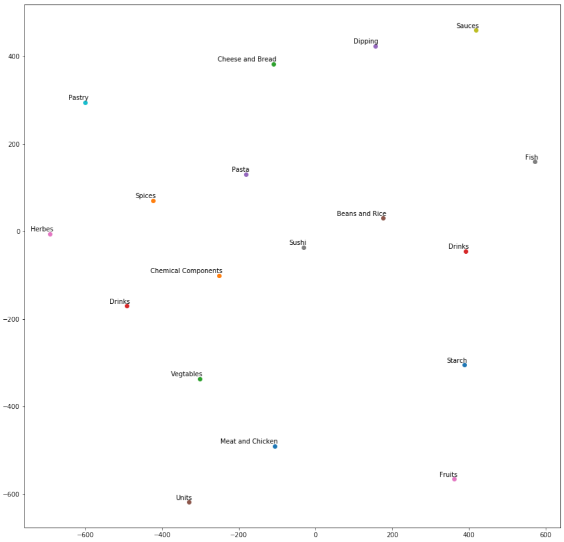

Figure 9. Ingredient clusters

5 Unified data set

The next step was to combine the cleaned and preprocessed data from the three websites into a single unified data set, and prepare the attributes so they can be efficiently analyzed.

5.1 Missing ratings

Since ratings had special importance for the study, keeping the values on a scale from 1 to 5 allowed more fine grained evaluation rather than compressed ratings on a 1 to 3 scale which was used for the preliminary analysis.

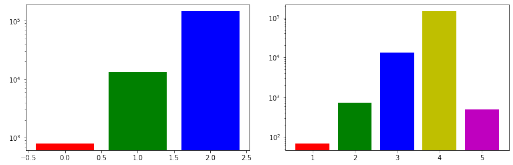

Figure 10. Rating distribution on scales of 3 and 5

The food-net site contained no rating values at all, only reviews. In order to retain this data, the missing rating values had to be enriched based on the reviews. This has been achieved by using sentiment analysis and logistic regression where both reviews and ratings were available, and use the trained model to predict the missing rating values from the corresponding reviews.

5.2 Rating bias

Rating values of 4 were by far the most common, resulting in an implicit bias in the analysis. In order to reduce this bias, the ratings have been slightly adjusted based on the reviews, resulting in a more balanced distribution. After all, the overall ratings provide a quick way to compare recipes, but readers are likely to read the corresponding reviews before deciding between them.

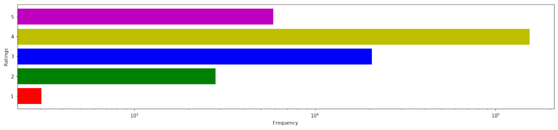

Figure 11. Improved rating balance

6 Analysis