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 is 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 is based on data extracted from the most popular websites containing food recipes. The data included recipes from over 50 websites, all featuring their own format and attributes. The initial analysis showed that a few websites containing the highest number of recipes represent a good portion of the total recipes. Since each website requires 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 give a good representation of the entire data set.

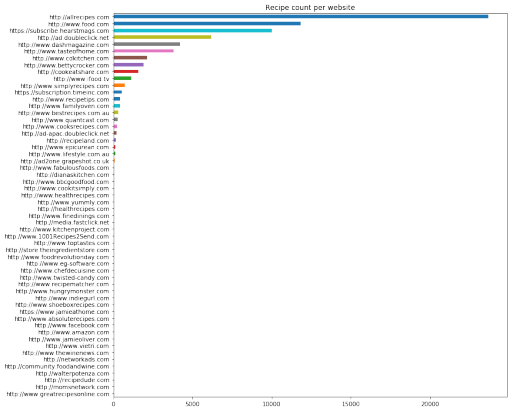

Figure 1. Recipe count

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

2.2 Data import

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

- 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 require custom scraping algorithms. The source data volume is about 13 GB, which is feasible to be processed on a single computer. The scraping takes 20-30 minutes, the results are stored in *json* format to allow efficient further processing.

2.3 Data cleaning

Due to the different formats, the data sets must be cleaned separately. Nutrition information has to be converted to numeric values, using the same measurement units across all recipes. Time values expressed in hours or days need to be converted to minutes. Rating values must be scaled to a common range. Some websites display servings as text, which must be converted to numeric values.

2.4 Review pre-processing

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

3 Preliminary visual analysis

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

3.1 Ratings

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

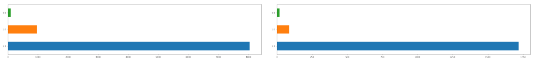

Figure 2. Ratings distribution

3.2 Serving size

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

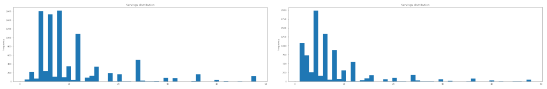

Figure 3. Servings distribution

3.3 Reviews over time

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

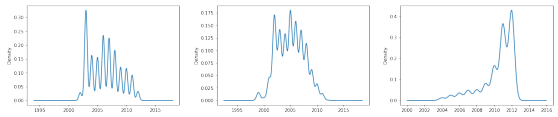

Figure 4. Reviews over time

3.4 Cooking time

Cooking times range for a few minutes to several hours, but the distribution shows that most recipes require less than 1 hour cooking, recipes requiring 30 minutes or less being more common.

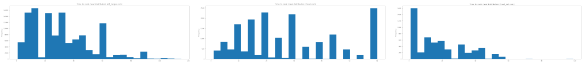

Figure 5. Cooking time distribution