Fraunhofer
PORTUGAL

RESEARCH CENTER FOR ASSISTIVE INFORMATION AND COMMUNICATION SOLUTIONS

1 *Image source: https://www.pexels.com/photo/adult-blond-board-brunette-298926.*

INFORMATION EXTRACTION FROM UNSTRUCTURED RECIPE DATA

An information extraction system for the food recipe domain

Motivation

Food recipes are an essential part of the lives of many individuals, who use these as a source of information for learning how to cook new dishes or as an aid for their food choice. Multiple websites on the Internet offer thousands of food recipes submitted by its users, which contain fields of structured information such as the name, ingredients, and directions, that allow for its users to filter through them according to their personal needs.

However, the structured information available for these recipes is often missing relevant information for its users, which can include the nutritional values of the recipe, cooking utensils required or each ingredient's applied cooking method. This information is often present in the recipe, albeit in an unstructured form. Finding a way to automatically retrieve and structure this information would allow for more fine-tuned searching and to improve the recommendation systems used by websites that offer food recipes. Being able to ea accurately determine a recipe's nutritional values using the extracted information could also help bring further clarity to the recipe's users on its nutritional content and overall effect on health.

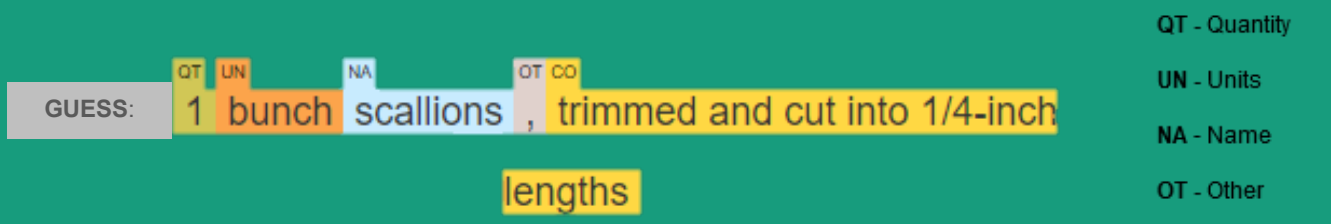
Goals

The objective of this project was the **development of a system that accomplishes the following goals:**

- The extraction of the name, quantity, applied cooking method and food preparation techniques of each ingredient in a food recipe;
- The extraction of the used cooking utensils in a recipe;
- The calculation of the nutritional values of a recipe, using the aforementioned extracted information in conjunction with a food composition database.

FRAUNHOFER AICOS
Rua Alfredo Allen, 455 / 460
4200-135 Porto, Portugal

+351 220 430 300
info@fraunhofer.pt
www.fraunhofer.pt



2

Segmentation

⇩

Classification

⇩

Association

⇩

Normalization

System

The system was developed in the Python programming language. Like any system performing an information extraction activity, it is divided into **four major tasks**:

- **Segmentation** of the recipe's text;
- **Classification** of the entities in the text;
- **Association** of the entities extracted;
- **Normalization** of the units and entities extracted.

Each of the tasks that compose the system has its own set of algorithms implemented, which range from simpler rule-based approaches to machine-learning ones. For this project, the algorithms for the extraction of the used cooking methods, food preparation techniques and cooking utensils were developed from the ground up, alongside the algorithms that associate ingredients to their respective applied cooking methods and food preparation techniques.

In addition to the information extraction activity task performed, the system calculates the recipe's nutritional values through the use of the extracted information, in conjunction with a food composition database.

Results

A set of 100 annotated recipes was used for testing.

For the extraction of cooking actions and utensils used in a food recipe, the system achieved an average F-measure of 0.89.

For the association of ingredients to their applied cooking method and food preparation techniques, the system achieved an average F-measure of 0.84.

For twenty ingredients with validated extracted information, the system was able to correctly associate eight ingredients to their database entries using the extracted information, an improvement over the three correct associations achieved by the baseline used.

The results suggest the system can reliably extract relevant information and associations in food recipes. They also imply it is possible to determine more accurate nutritional information for each ingredient through the use of additional structured information.

2 *Extracted entities from an ingredient line.*