TO BE IN SEASON OR NOT TO BE IN SEASON

EKATERINA DOPIRO, FABIEN ZELLWEGER & CHRISTOPHER BENZ

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

DATA SCRAPING:

Introduction

Humankind has acquired knowledge about food for ages. As a result we know that each food has its season. However with the rise of civilization we are no longer dependent on it as any food can be transported to any point in the world.
The question is then: how much food do we consume out of the season? Our project focuses on two aspects: the season the foods are naturally produced, and the area where they are naturally produced.

Data Scraping

The first challenge of our project is to get a lot of recipes including their ingredient list and the review dates. We also need the seasons of those ingredients as well as their states of origin.

- Extract review date from our 3 most present websites (allrecipes.com, food.com and foodnetwork.com).
- Extract the corresponding ingredient list on our dataset.
- Extract month per ingredient from different source, including, if possible, the states (eattheseasons.com, seasonalfoodguide.org).

Percent of website in our dataset

http://allrecipes.com/	25.66 %
http://www.food.com/	13.27 %
http://www.foodnetwork.com/	10.86 %
http://www.yummly.com/	5.96 %
http://www.cooks.com/	5.02 %
http://www.epicurious.com/	4.57 %
http://www.tasteofhome.com/	4.36 %
http://www.myrecipes.com/	3.44 %
http://recipes.sparkpeople.com/	3.32 %
http://www.cdkitchen.com/	2.34 %

Figure 1: The proportion of the 10 biggest websites in our dataset. Given this distibution, we choose to only take care of the top 3.

DATA PROCESSING

Scoring

We introduce a score in order to rank the food consumption as seasonal or not. For a given food, a score is computed for each month of the year: the score is equal to 6 minus the number of months to the closest month when that food is consumed.

$$S_i \in [0, 6], i \in [1, 12] \tag{1}$$

$$S_i = 6 - [\#\text{months to closest seasonal month}] \tag{2}$$

Estimates the distance in time and place to the optimal climate for the production of the given food. Then, for the recipe score, we take the mean food score on all foods in the recipe.

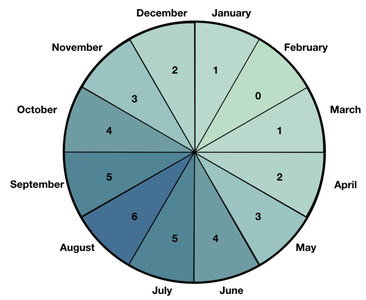


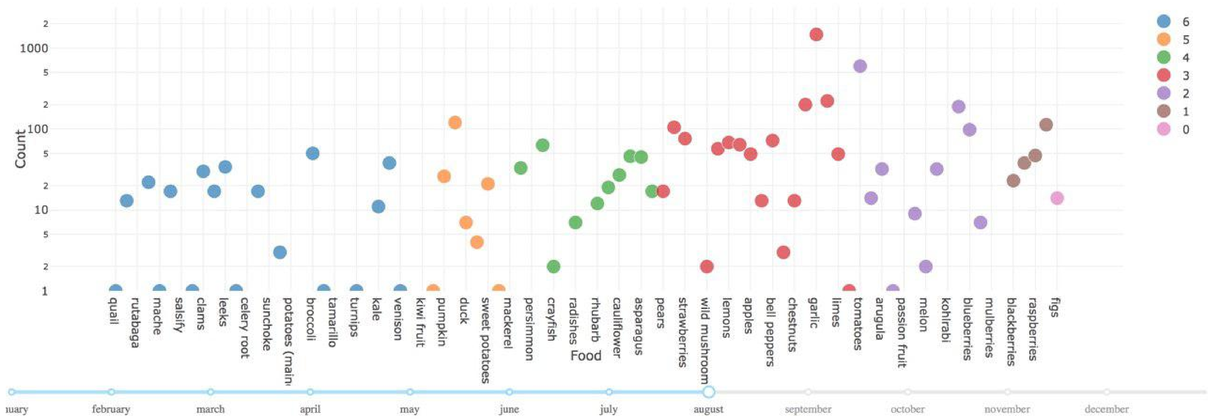
Figure 2: The score for a food which is seasonal on August

Words matching

The list of ingredients for a recipe contains a lot of noisy data, e.g. in 'one and a half tablespoons of black pepper', we are only interested in 'black pepper'. Therefore, for matching between the two datasets, we search for any occurence of the food name in the recipe ingredient list. This causes issues for double-worded foods (e.g. 'sweet potato') because it matches both on 'sweet potato' and 'potato'.

RESULTS

Results



Our visualization results allow us to observe some facts about the data. For example: - Garlic seems to be highly consumed no matter the season, compared to other foods. This could be explained by the fact that it is usually used as a spice, and that people love garlic bread all year long. - Consumption at the correct season seems to dominate during the late winter and spring months, while during summer and autumn there is much more un-seasonal consumption.

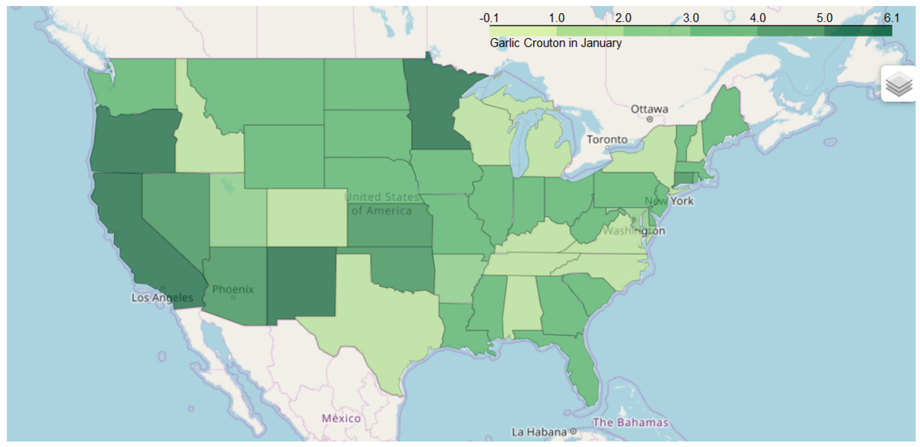


Figure 3: The score for the Garlic Croutons recipe for January

Conclusion

To conclude, our brief consumption analysis has shown that overall we are pretty good at consuming the foods when we are supposed to, but there is room to improve, especially in months of summer.