Machine Learning for Building a Food Recommendation System

Luís Rita[1]

Supervision: Kirill Veselkov[1], Michael Bronstein[2]

(1) Department of Metabolism, Digestion and Reproduction, Imperial College London

(2) Department of Computing, Imperial College London

Aim

Use the largest publicly available collection of recipe data to build a recommendation system for ingredients and recipes. Train, evaluate and test a model able to predict cuisines from ingredients. Estimate the probability of negative recipe – drug interactions based on the predicted cuisine. Finally, to build a web application as a step forward in building a 3D recommendation system*

* Considers user preferences, positive and negative compounds

Introduction

WHO Europe — Cancer is 2nd leading cause of death globally [1]

Cancer Research UK — Healthier diets could prevent around 1 in 20 cancers [2]

WebMD — Body needs calories and nutrients to stay strong. Disease can make it hard to get enough, which can be different before, during, and after treatment [3]

Springer — Studies on the food domain focus on recommendations that suggest food for each user considering preferences and health problems [4]

Methods

Recipe1M+

[Largest publicly available collection of recipe data. Ingredients not separated from text. Use of vocabulary to extract them]

[{"ingredients": [{"text": "1/2 cup green onions, chopped"}, ...], ..., "title": "Salmon & Salad a La SPORTZ"}, ...]

Kaggle and Nature

[Dataset to train cuisine classifier. Recipes with the (tokenized) ingredients and cuisines available]

African, honey, yeast, wheat

EastAsian, peanut_oil, oyster, wheat

LatinAmerican, lard, vegetable_oil, wheat

Pre-Processing

[Ordering ingredients alphabetically to make context independent of initial position]

honey yeast wheat

peanut oil oyster wheat

lard vegetable oil wheat

Word2Vec

[Representing ingredients in 100-dimensional vectors]

honey: [0.3 - 0.7 ... - 2.3]

yeast: [2.1 0.3 ... - 2.0]

wheat: [- 0.4 - 1.4 ... 1.9]

Recipe to Vector

[Averaging all ingredients' vectors in the recipe]

[0.7 - 0.6 ... - 0.8]

[1.1 0.2 ... - 2.3]

[1.6 - 1.1 ... - 2.0]

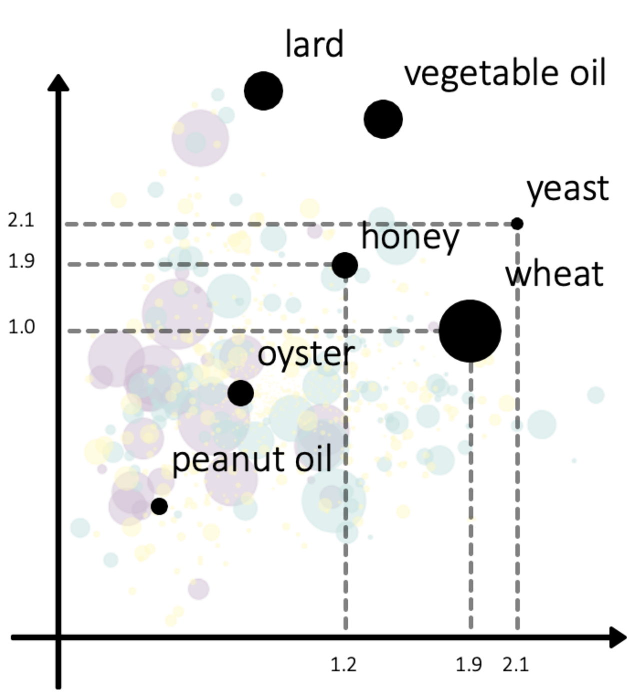


Figure 1 Colours represent clusters of ingredients used more often together. Sizes are proportional to the number of occurrences in the datasets

PCA

[Reducing vector dimensions from 100 to 2]

honey: [1.2 1.9]

yeast: [2.1 2.1]

wheat: [1.9 1.0]

SVC

[Training classifier to predict cuisines from list of ingredients]

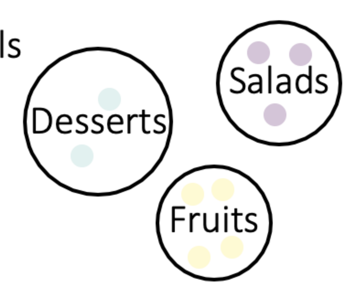
African

East Asian

Latin American

Spectral Clustering

[Grouping ingredients in 9 clusters accordingly to their similarity]



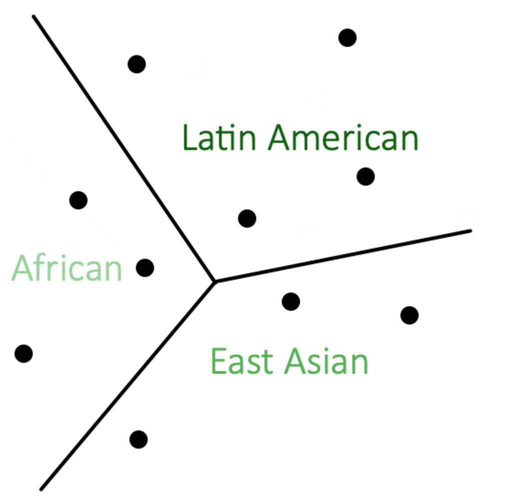


Figure 2 Linear kernel SVC used to train the model. Confusion matrix built to evaluate the accuracy. Tests were performed

Web

[Backend Node.js server and a static frontend were implemented]

Python Shield	Load Word2Vec and SVM models
Express	Connect to frontend
HTML	Website block structure
CSS	Style personalization
JavaScript	Connect to server and receive user input

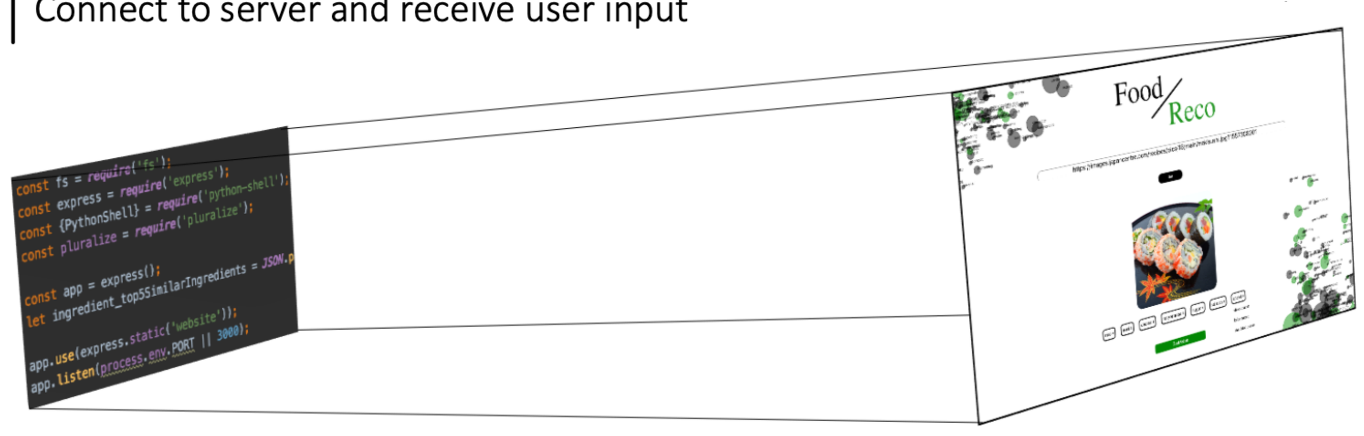


Figure 3 Tools detailed in Methodology used to create a food recommendation web application

Results & Discussion

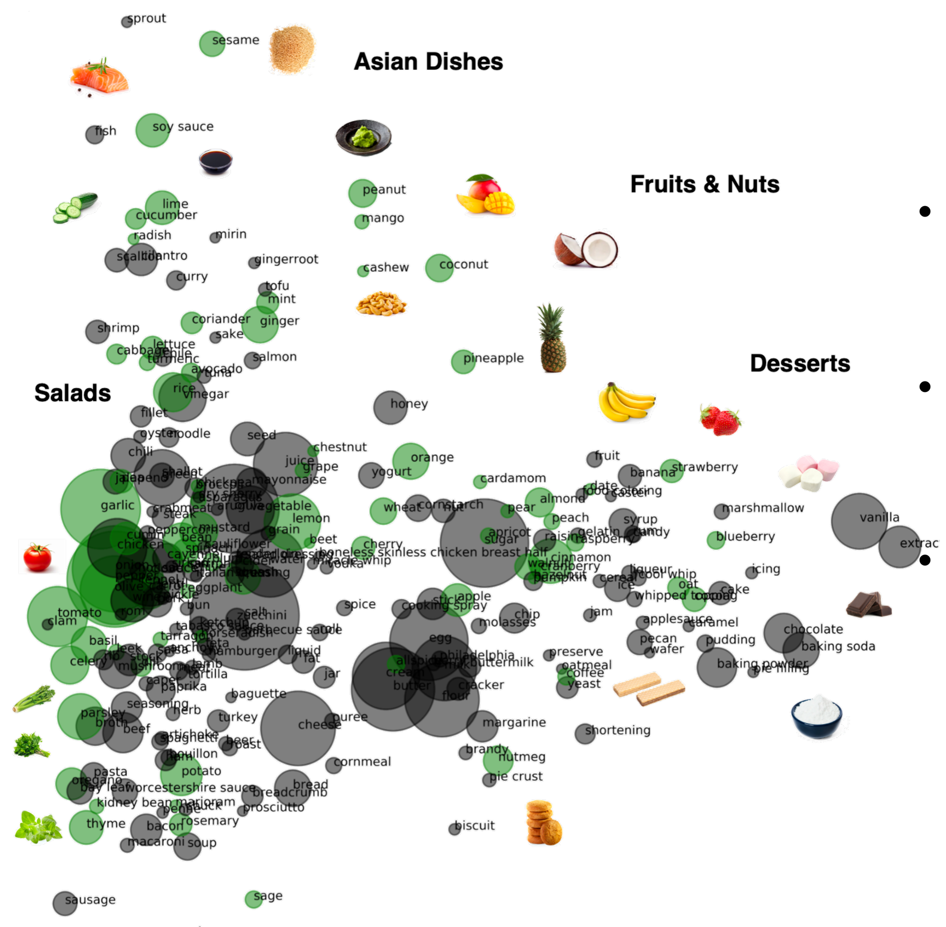


Figure 4 Ingredients occurring > 800. In case of green nodes, they contain, at least, 1 cancer-beating molecule

Ingredient Recommendation

- Successful food combinations closer in the plot. Confidence increases with the size of the nodes
- Spectral clustering identified *Salads*, *Asian Dishes*, *Fruits & Nuts* and *Desserts* categories
- Colours allow to account for the presence of cancer-beating properties

Cinnamon | Walnut | Cranberry | Peanut | Mango

Model Evaluation

- NE, EE, WE and ME cuisines often misclassified as NA
- European cuisines often misclassified among them – share of ingredients and under representation in the dataset may be some of the reasons
- Highest accuracy achieved for EA, SA and NA

Figure 5 Confusion matrix

	NA	WE	NE	EE	SE	ME	SA	SEA	EA	LA	A	Total
K&N	45843	6774	739	381	14178	645	3618	3572	7435	11892	1173	96250

Table 1 Cuisines distribution in K&N

Model Test

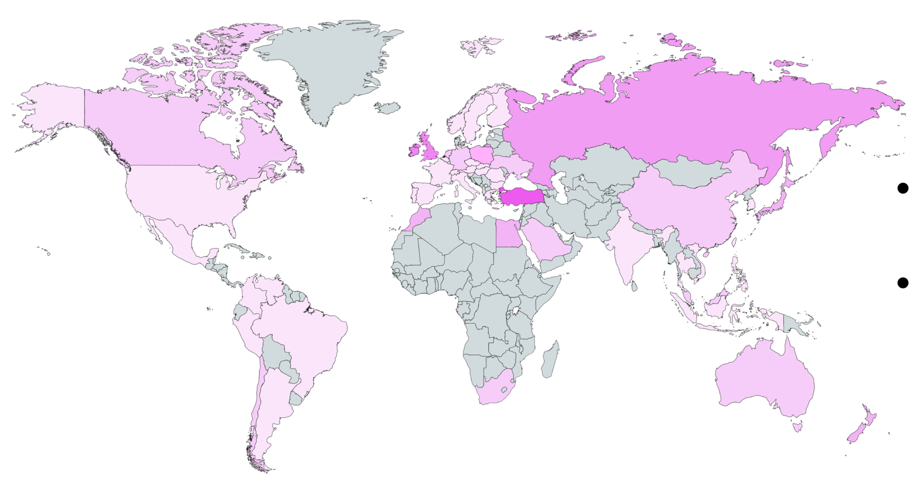


Figure 6 Tea consumption per year per person [5]

- Predicted cuisines for every recipe in Recipe1M+
- All recipes with word *tea* in title were retrieved. Calculated and compared cuisines distribution with initial. NE and EA presented higher variation

	NA	WE	NE	EE	SE	ME	SA	SEA	EA	LA	A	Total
Salad	20%	1%	6%	3%	36%	1%	1%	6%	9%	14%	3%	62929
Tea	12%	6%	25%	1%	5%	1%	5%	5%	27%	3%	8%	4212
Recipe1M+	14%	4%	16%	2%	21%	1%	2%	13%	10%	15%	3%	1029356

Table 2 Cuisines distribution in Recipe1M+. Plus, filtering the ones containing keywords *salad* and *tea* in title

Cuisine Classification

- Calculated average presence of cancer-beating molecules for recipes across cuisines
- Top classified: Mediterranean diet

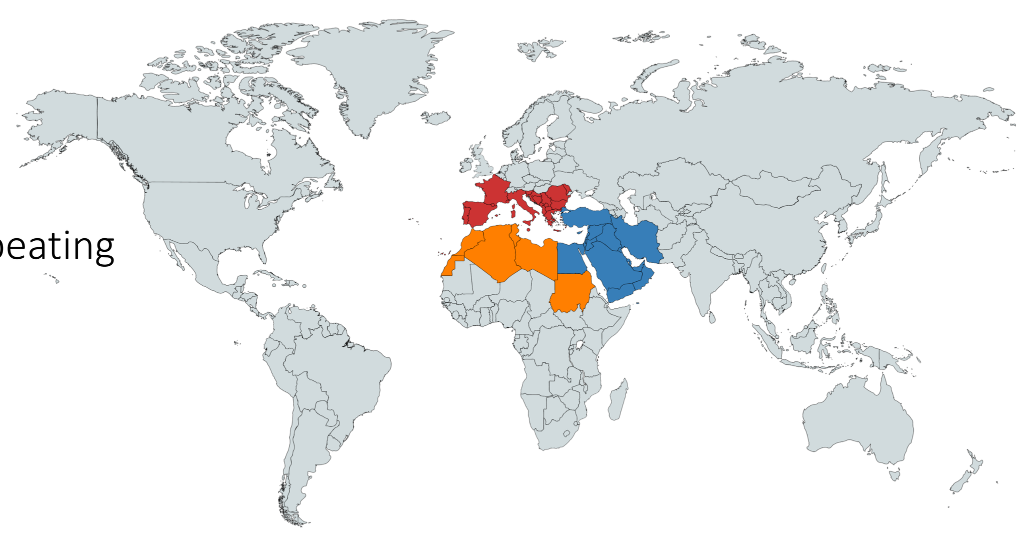


Figure 7 Cuisines with higher number of cancer beating molecules per recipe

	NA	WE	NE	EE	SE	ME	SA	SEA	EA	LA	A
Anticancer Score	1.25	1.49	1.30	1.90	2.21	2.08	1.90	1.25	1.32	1.75	2.48

Table 3 Average number of cancer beating molecules per recipe in each cuisine

FoodReco

Food / Reco

Figure 8 Web app logo

- Receives an image URL. Outputs predicted list of ingredients and cuisine
- Top 3 best substitutions are suggested for each ingredient
- Indication on the expected number of negative interactions with antineoplastic drugs

References

[1] "Cancer," World Health Organization, 12 September 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed 7 March 2020]

[2] "Diet and cancer," Cancer Research UK, [Online]. Available: https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/diet-and-cancer. [Accessed 7 March 2020]

[3] "How to Eat When You Have Cancer," WebMD, [Online]. Available: https://www.webmd.com/cancer/cancer-diet#1. [Accessed 7 March 2020]

[4] T. Tran, M. Atas, A. Felfernig and M. Stettinger, "An overview of recommender systems in the healthy food domain," *Journal of Intelligent Information Systems,* vol. 50, p. 501–526, 2018

[5] "Annual per capita tea consumption worldwide as of 2016, by leading countries" statista, 14 January 2016. [Online]. Available: https://www.statista.com/statistics/507950/global-per-capita-tea-consumption-by-country/. [Accessed 7 March 2020]