Recipe explorer

Axel Vandebrouck
Msc in SC, EPFL Lausanne

Quentin Rebjock
Msc in DS, EPFL Lausanne

Bojana Ranković
Msc in CS, EPFL Lausanne

***Abstract*—While starvation has almost disappeared in most developed countries, people are nowadays more likely to die from overeating than from a lack of food. We eat badly and it's a serious public health problem in our modern societies. Even if taking the healthiness aspect in food recommendation systems won't entirely solve the problem, it can help people without special knowledge to eat the right food. These recommendations should take into account the user's tastes and give further information about the nutritive aspect of the recipes.**

I. Introduction

The 21st century represents one of the most progressive eras in human history. We are busier than ever. We live fast and we eat fast. Our goal is to create a recipe recommendation system that would allow users to enjoy generated meals according to their preferences, physical activity and taste while allowing them to see information about the healthiness of their choices.

To show the approximated healthiness of the recommended recipes, we have considered the proposed nutritional requirements recommended by FDA[1]. Based on this, a healthy menu for an adult person consists of meals that fill the 2000 kcal daily goal.

The last part of our project takes into consideration that cooking is not only about aiming to fulfill some proposed values in order to be healthy. Beside that, it is usually fun. By answering the question about cuisine and ingredients similarities, we offer our user to explore the world of food in a way which will hopefully make it more interesting to follow a certain healthy diet.

II. Related work

Recommendation systems have been widely used with variety of topics, but their popularity in recent times comes from the ever-growing online database about users. However recipe recommendation systems, given their cause to provide people a good meal recommendation, often face a challenging task of understanding the overall health benefits of a certain food. The most recent work regarding this subject can be found in [1] Also, combining different recipes into a menu for the purpose of building meal planner would involve different optimization tactics like shown in [3]. The food research, however, goes beyond just recipe recommendations and involves exploration of different eating habits across the world. [4]

III. The recipes dataset

A. Data collection

The Internet has become a source of non exhaustive information regarding our daily lives. The challenge lies in obtaining the data that would best suit our needs. Even though we did not have the appropriate data set provided upfront, we explored the storage of Internet in order to use the unlimited knowledge about recipes available online. We have performed scraping of one of the most popular food inspired search engines, yummly.com. The initial data set is available online as *Yummly-28K* and contains around 28 thousand recipes in json format. However, for the goal of building recipe recommendation system, we needed to gather data on users. By scraping the web page of each recipe in the initial data set, we collected 5037 reviews from 4328 users, which gave us a baseline for creating recommendation system. However, it was later noticed that the data obtained in this way is rather sparse, so the performance of our recommendation system would not be satisfying. As a result of the structure of the *yummly* website we were able to overcome this problem by downloading data on liked (saved) recipes for each of our previously collected users. This in the end gave us enough data to build a valid recommendation system.

B. Data description

The collected data has been then split into three major *.csv* files :

- One containing the recipes ratings and the user saved recipes. It was mainly used to create the recommendation system.
- One containing detailed information about recipes, including cuisine type, course type and ingredients.
- And finally one with the nutritive values of each recipe to perform the healthiness study of the recipes.

In overall, we explored $33,872$ recipes from 33 different cuisines. The number of users we considered is $4,046$ that saved a total of $109,264$ recipes as their favourite meals on the yummly website, which gave us an average of 27 recipes per user to base our recommendations on.

IV. Data analysis

We are presenting in this part the approach that was followed to analyze the collected data and what decisions were taken.

A. Food healthiness investigation

As lifestyle-illness like diabetes and obesity reach more and more people, especially in the USA, it seemed essential to take the healthiness of the recipes into account in our recommendations. This epidemic reduces life expectancy and costs a lot to health care systems but is actually pretty easy to stop: it is all about learning how to eat properly.

[1] https://www.fda.gov/AboutFDA/Transparency/Basics/ucm194877.htm

The goal was to get an user-friendly way to evaluate how healthy some food is and visualize it in a way such that anyone can understand it without having advanced notions of nutrition. There are several ways to measure the healthiness of a recipe and none of them is perfect. A first step can be to compare the quantity of different nutrients to the daily value references (DVR). However, raw percentages are usually hard to interpret for inexperienced people and values corresponding to little known nutrients like selenium is pretty irrelevant to the consumer.

That is why the traffic light rating system was implemented instead. It consists in assigning a score to each of the attributes **total fat**, **saturated fatty acids**, **sugar** and **sodium** (salt) which are the main nutrients and indicators of the healthiness of the food. The range of the scores is 1 (good) to 3 (bad) and they are then summed to get a global score for the whole recipe. However, since the recipes contain different quantity of food, it is unfair to compare the raw values of these nutrients instead, they are normalized according to the weight of the recipe.

The scores obtained are thus between 4 and 12, the range 4-6 indicating that the food is quite healthy, the range 7-9 being average, and the range 10-12 corresponding to pretty unhealthy recipes (to consume with moderation).

B. Recommendation system

Using all this data, we first built a basic recommendation system which recommends recipes to some users comparing the recipes they saved with other users. This is called user-item recommendation. First, we build a matrix Sim of similarities between recipes using cosine similarities. More specifically, this matrix will contain at position $(r1, r2)$ the cosine similarity between recipes $r1$ and $r2$.
Once this is done, we look at the users known saved recipes and their similarities. Basically, we create a score for each recipe we have for that user and then we can simply choose the recipes with the highest score. The score attributed to recipe i for user u is computed as follows:

$$s(u, i) = \frac{\sum_{j \in \mathcal{R}} Sim(i, j) r_{ui}}{\sum_{j \in \mathcal{R}} Sim(i, j)}$$

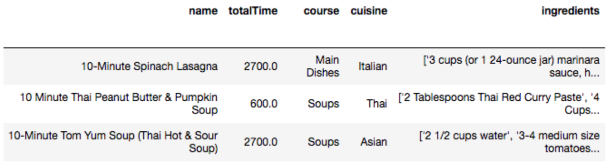
where $Sim(i, j)$ denotes the similarity between recipes i and j, r_{ui} is an indicator function which takes the value 1 if user u has saved the recipe i (and 0 otherwise), and $\mathcal{R}$ denotes the set of recipes.

Once we have done this, we use this recommendation system to create a more advanced and interactive one which works as follows: The user selects the constraints from which he wants his recommendations. These constraints can be of different types. The first one is nutritional, whereas the second one indicates the maximum time of cooking. Then he can select the type of cuisine and course he wants. Finally, the recommendation system filters the recipes with respect to those constraints and then computes the score of the remaining recipes to propose the n best ones in real-time, where n can also be chosen by the user. Furthermore it displays the nutritional values of those recipes and also gives links to the recipes web pages.

C. Recipe explorer

One of our research question was whether we could infer data patterns in recipes. Do ingredients tell us something about the cuisine from which the specific meals comes from? Is there a connection between ingredients that can be discovered by observing their frequency in specific types of meal, from specific cuisines? The last part of our analysis is dedicated to discovering these patterns.

In order to answer these questions we use the data set containing the data on recipes from various cuisines. Each recipe comes with a list of ingredients so the cuisine similarities and the ingredient embedding are deducted from it.



name	totalTime	course	cuisine	ingredients
10-Minute Spinach Lasagna	2700.0	Main Dishes	Italian	['3 cups (or 1 24-ounce jar) marinara sauce, h...
10 Minute Thai Peanut Butter & Pumpkin Soup	600.0	Soups	Thai	['2 Tablespoons Thai Red Curry Paste', '4 Cups...
10-Minute Tom Yum Soup (Thai Hot & Sour Soup)	2700.0	Soups	Asian	['2 1/2 cups water', '3-4 medium size tomatoes...

Fig. 1. Recipes data sample

Since this is the data set which would give answers about different questions on cuisines, it is important to know how well the cuisines are distributed. The figure 2 shows that our data is skewed, with a majority of American recipes, followed by Italian and Mexican cuisine.

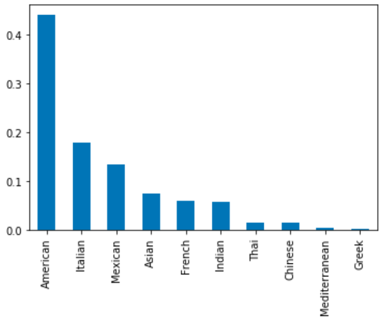
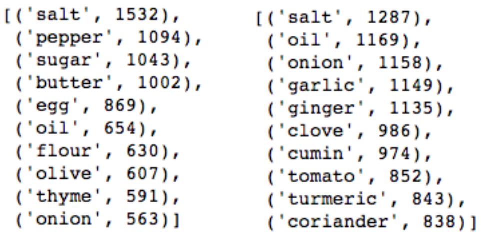


Fig. 2. Cuisine distribution

1) Cuisine similarity analysis: For the goal of calculating cuisine similarities, we needed to explore the ingredients data for each recipe. As our data set comes from the real world, where the data can be ”messy”, it was required to perform cleaning, by removing the numerical values for recipe, and special stop words used in kitchen, like measure units and the description of ingredients. The cuisine similarity would then be possible to construct by taking the cosine similarity between the vectors of two cuisines. The cuisine vector in this case is constructed by taking the ingredients found in recipes of that cuisine and calculating their frequency in the cuisine. The example can be seen in 3. After calculating the vectors for each national cuisine in our data set, we can calculate cosine



```
[('salt', 1532),
 ('pepper', 1094),
 ('sugar', 1043),
 ('butter', 1002),
 ('egg', 869),
 ('oil', 654),
 ('flour', 630),
 ('olive', 607),
 ('thyme', 591),
 ('onion', 563)]
```

(a) French vector

```
[('salt', 1287),
 ('oil', 1169),
 ('onion', 1158),
 ('garlic', 1149),
 ('ginger', 1135),
 ('clove', 986),
 ('cumin', 974),
 ('tomato', 852),
 ('turmeric', 843),
 ('coriander', 838)]
```

(b) Indian vector

Fig. 3. Example of vectors for calculating cuisine similarity

distance with: $cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$ where x and y are our cuisine vectors.

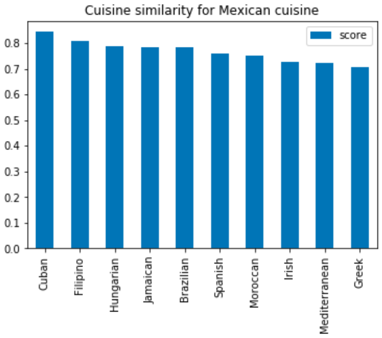


Fig. 4. 10 most similar cuisines to Mexican cuisine

It proves interesting to plot these similarities onto the world map. This could provide better overview on the question whether geographically close countries share the similar taste for food. For the visualization of the similarities, we map the cuisine names to different countries or regions, which gives us an overview of world cuisines through a similarity map.

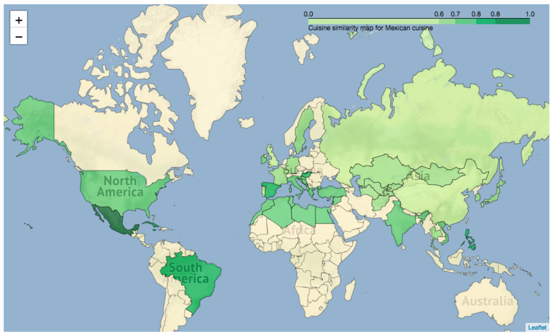


Fig. 5. Similarity map for Mexican cuisine

2) Ingredient embedding analysis: We have seen how connection between recipes, ingredients and their originated cuisines provides an overview on how close a national cuisine can be to another. But the next question we wanted to explore, should provide an answer to how ingredients themselves form strong connections in different cuisines. The main idea for this part is to use word embeddings on recipe ingredients. Word2Vec is a group of models that use shallow two-layer neural networks to reconstruct linguistic contexts of words and produce word vector. These vectors hold the property to be located in close proximity to one another in the vector space if their words share common contexts in the corpus.

First we needed to check the ingredients number distribution for each cuisine to be sure that the differences in calculated vector would not be the result of inconsistencies in data, but solely the attributes of cuisines. The majority of recipes observed share the similar number of ingredients as observed from figure 6.

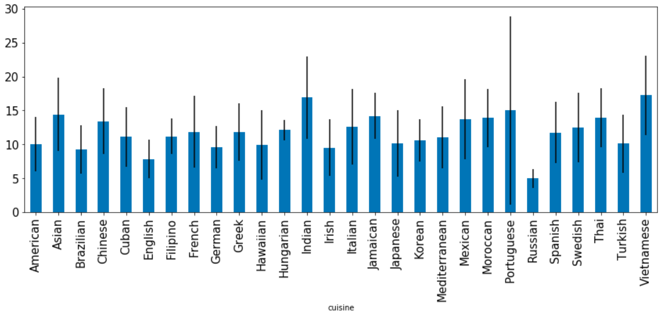


Fig. 6. Ingredients number distribution per cuisine

The most recipes in our data set have an average number of ingredients around 12. This is set as a context window for the word2vec algorithm, while the word vector dimensionality takes the number of approximately 50% of the ingredients corpus, which is around 1500. Some of the ingredients may be omitted if their frequency is below the threshold of 5 in the whole corpus. With these parameters our model shows good understanding of the ingredients. The most similar ingredient to pasta is spaghetti, while cheese goes with macaroni in most of our recipes, which may be due to fact that the biggest portion of recipes comes from American cuisine where ”Mac&Cheese” is one of the favourite meals. In order to visualize our ingredient vectors in two dimensional space, we use the ’tSNE’, a technique which is a variation of Stochastic Neighbor Embedding, proved to give good results in dimensionality reduction. [2]

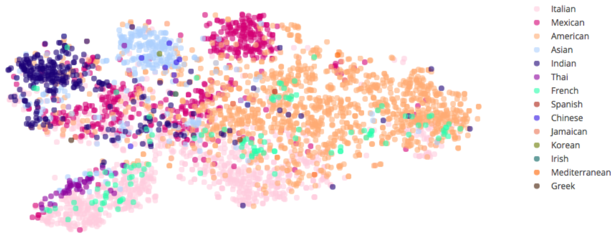


Fig. 7. Ingredient embeddings

V. Results and findings

In regards to the healthiness of the collected recipes, it turns out that they are overall quite unhealthy. We could expect that because these meals are supposed to be hearty and not made to be low in calories. Among all the recipes, we kept only those that weigh than 100 grams for the analysis because there are some recipes in the data set which would not count for a full meal. Out of these 12,179 recipes, the majority of them gets a red flag (around 63%) and they are in average unhealthy : the mean is around 9.74 and only 1% are green, and 35% orange. Here is the distribution of the scores to get a deeper insight:

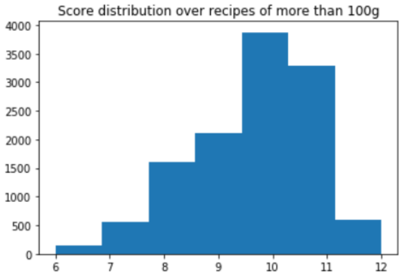


Fig. 8. Score distribution over recipes of more than 100g

The following box-plot is more easily interpretable and clearly shows how unhealthy the food is overall:

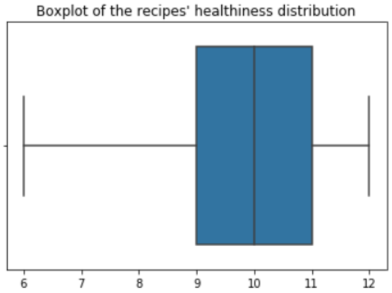


Fig. 9. Score distribution over recipes of more than 100g

The median being almost 10 (red flag), it appears distinctly that one must be cautious when eating such recipes !

For each recipe, we would like the user to be able to visualize easily the scores that were computed. A very intuitive way to do so is to match a color to each score : green, orange and red for respectively healthy, moderately healthy and unhealthy. Here is an example of the kind of plots we get :

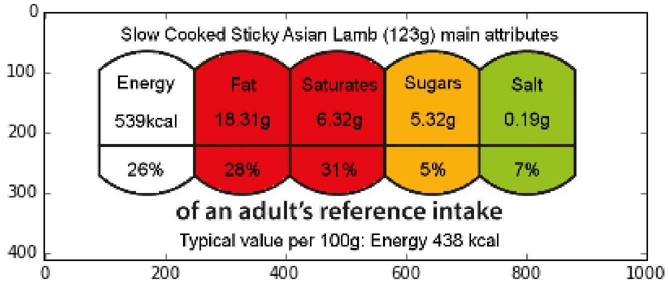


Fig. 10. Example of user-friendly traffic light plot

From that kind of labels, a consumer can immediately see how healthy the food is even without having advanced nutrition knowledge.

The cuisine explorer provided a set of interesting answers to our research questions. Indeed, there is a high correlation between ingredients and the cuisine in which they are most commonly used, which was observed from the clusters in our ingredient embeddings map. Transforming ingredients into vectors gave us an overview of how people use ingredients in their culinary habits. The result of our word2vec transformation shows that the most similar ingredients to pasta are spaghetti, penne, while the ingredients close to pork is beef. Another interesting result we extracted from our model is that the bacon is to pork as spaghetti to pasta, so we conclude that the model created based on recipe ingredients learned the world of culinary quite well. We have observed interesting clusters of ingredients, as a result of the word to vector implementation. This can set a path to more interesting questions. Can we calculate if a specific cuisine is a mixture of other world cuisines? What would be created when mixing recipes from different cuisines? Can we find substitute ingredients based on the ingredients?

VI. Conclusions

Finally, the recommendation system that we developed is giving reasonable suggestions taking into account the healthiness of the food. The users can still ignore that last part, but we hope that seeing these colored labels for nutrients will encourage them to make the right choice. As most of the recipes are pretty unhealthy, it is hard to find the right balance between recipes the user likes and healthy ones.

With regard to a possible continuation of the work done, the very first step would be to make the interactivity easier, having a website dedicated to that instead of using a notebook. Also, the recommendation system could be improved using several techniques, among other things by combining several types of recommendation systems. Ultimately, the idea would be that the data of each user is saved in real time, including other aspects than just the recipes they liked, and recommend full menus considering the life-style of the user. Another enhancement much harder to implement would consist in automatically adjust and modify the recipes so they keep tasting and looking the same but being more healthy.

VII. References

References

[1] Christoph Trattner, David Elsweiler, and Simon Howard. *Estimating the Healthiness of Internet Recipes: A Cross-sectional Study*, volume 5. Frontiers Media S.A., 2017.

[2] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[3] Longqi Yang, Andy Hsieh, Hongjian Yang, John P. Pollak, Nicola Dell, Serge Belongie, Curtis Cole, and Deborah Estrin. Yum-me: A personalized nutrient-based meal recommender system. 36:1–31, 07 2017.

[4] Longqi Yang, Andy Hsieh, Hongjian Yang, John P. Pollak, Nicola Dell, Serge Belongie, Curtis Cole, and Deborah Estrin. Yum-me: A personalized nutrient-based meal recommender system. 36:1–31, 07 2017.