A Visual Approach Towards Food-Ingredients Ranking & Classification

TEAM #125 – PROGRESS REPORT

Alessio Van Keulen | Fei Hui | Liu Yuxin | Rutuja Patil | Sufian Suwarman

Introduction – One of the factors that influences purchase decisions for health aware consumers, is the ingredient list of processed food items. For example, a person following a Keto diet will try to avoid carbohydrate-rich products, and for that they will be filtering out those that list Gluten as main ingredient. However, scientific nomenclature along with nutritional information, can be confusing and hard to decipher, and will typically frustrate the buyer leading them toward making the wrong choices. Our project will aim at simplifying the purchasing process by providing the customer with a tool capable of giving them the information they need at-a-glance. Supported with the data, the end-users will be quickly able to draw their own conclusions as to whether the just-around-the-corner supermarket is a healthy shopping habit or not.

Innovative Methods – Food products will be classified based on their ingredient list, they will be ranked, scored, and presented through a visual interactive interface. Traditional methods of grocery store classification involve surveying a geographical area to gather insights and preferences, brand trustworthiness, mouth-to-mouth spread of information, comparative advertisement, etc. We will overcome the resources, biases, and logistic challenges associated with these practices with a modern, efficient, data-driven approach. Success will be evaluated by comparing the results of our algorithm with other scientific studies. Since the application will live on a website, acceptance, usage, and measures of popularity can be obtained by simple web analytics, at no-cost. *Figure 1* outlines our novel approach:

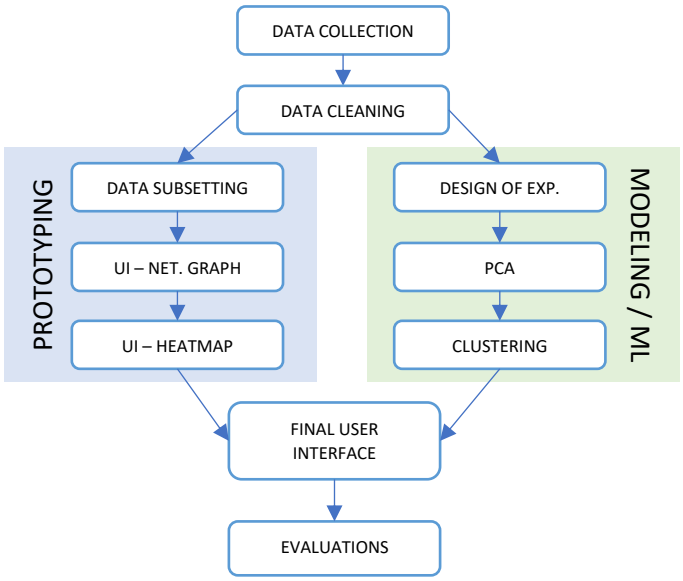


Figure 1 - Workflow

- Grocery store items data collection and cleaning
- Visual interface and interaction prototyping
- Modeling and Machine Learning algorithms
- Final build, evaluations, and measurements.

We believe that both the *Prototyping* and *Modeling* phases are the backbones that sustain the innovation aspect of the process. Especially in the prototyping phase, we will be pioneering user interface and interaction with the goal of maximizing the amount of information the user can gather in a short amount of time.

Detailed Procedures – Each of the above-mentioned processes is broken down in these steps:

1. Data Collection | The data sustaining this study comes from an open database (Kaggle [ǂ1]) and contains ~10K unique food listings each with their own set of ingredients. Each observation (row) represents a processed food product and related information spread over 14 attributes.

2. Data Cleaning | Among all the attributes we are mostly interested in the list of ingredients: “features.value” variable. These were listed using complex unstandardized formats, including non-

ingredient related information, such as allergy warnings, slogans, non-alphanumeric characters, etc. Regular Expression (regex) technology was used to parse and clean these entries. Then, the cleaned data was melted (figure 2) to transform it from a long structure to a wide structure: row containing unique food products with an associated list of ingredients, now contain repeated food products names for each of the ingredient found within. The melted dataset now contains roughly 62.5K observations. This format was crucial in building one of visual tools: heat map, covered later. Along with melting the data we've developed a Python algorithm capable of finding all possible combination of ingredient taken two a time for each food product. This data format introduces the next visualization tool: network graph of ingredients correlation.

Figure 2 - Data Melting

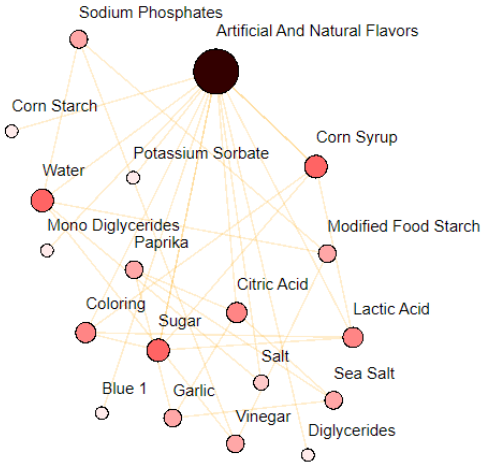


Figure 3 - Network Graph of Ingredients Correlation

3. Prototyping – Network-Graph |Working on a smaller chunk of the dataset we've prototyped our first visual interactive interface: Network Graph of ingredient correlation. Figure 3 shows a very minimal subset of the data, significant enough the demonstrate a basic network. Although this screenshot showcases < 1% of the original data, we can already identify a pattern: it would seem that the ingredient "Artificial and Natural Flavors" is connected to many of the other ingredients found in this chunk of data. This multiple connection is reinforced by the larger radius of the circle, as well as a darker saturation of the red hue chosen for this representation. Albeit premature, this simple illustration already carries a strong and suggestive message that a user could interpret as a negative score against that grocery store, source of the collected dataset. The network graph will effectively describe how "viral" an ingredient is in a set of food products, by displaying all its ramifications and links (edges) with other ingredients.

4. Prototyping – Heatmap | To provide the user with a more quantitative visual output, we've supplemented our interactive chart with a heatmap: Figure 4. Notice how the chose color is the same for the network graph: each tassel represents the frequency of an ingredient, and the saturation is the same as the related node. In fact, upon hovering the colored block, it will zoom-in and contemporary highlighted the related node in the network-graph.

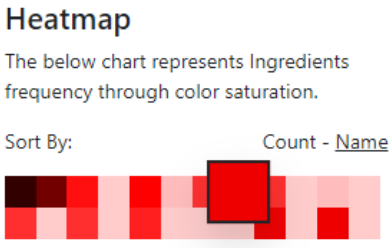


Figure 4 - Heatmap of Ingredient Frequency

5. Final User Interface | Lastly, to supply consumer with more detailed information about each food product that originated these two user interfaces, we've produced a secondary dataset consisting of a sample of 12 products and their associated information. This includes:
    - Ingredients,
    - Processing classification (according to NOVA classification system [ł2]) ranging from NOVA-1: unprocessed, to NOVA-4 ultra-processed food items.

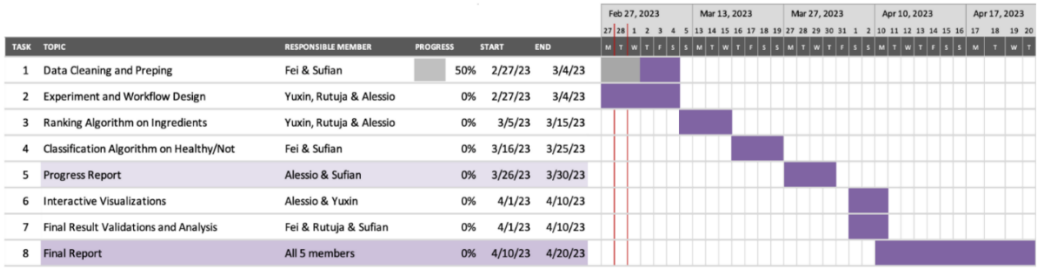
- Nutritional scores (according to Open Food Facts [ŧ3]) ranging from A to E with A being the healthiest and E the least healthy,
- Nutritional Value such as product calories and the quantities of fat, saturated fat, sugar, etc.

The food products are selected randomly 3 at a time for each of the NOVA classifications.

Design of Experiments

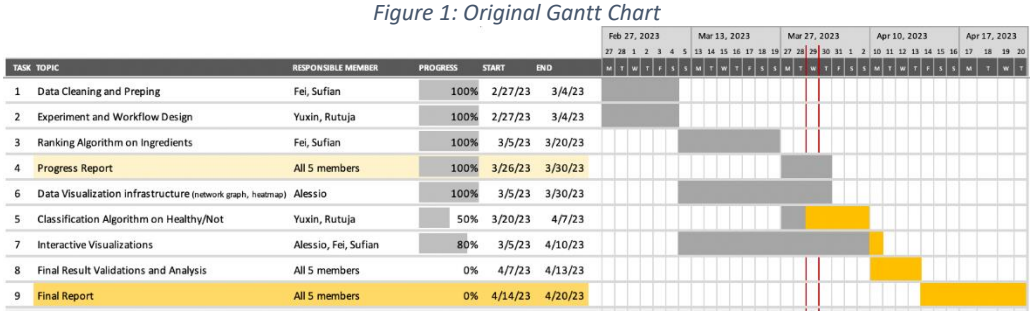
Experiments & Evaluation To be able to create our own ranking and categorization of food, our approach was a clustering algorithm due to the lack of explicitly labeled data. Our two approaches include a hierarchical clustering algorithm and a DBSCAN algorithm. Another variable selection step was undertaken to make sure correlated columns would be removed from the dataset and would not impact the accuracy of the models. The main advantages of applying this hierarchical clustering form are the ability to obtain detailed information about similar observations. Its two key abilities: lack of sensitivity to initial conditions and categorical variable inclusion are some of the key advantages as well which is why we chose to utilize the hierarchical algorithm in the first place.

Plan of Activities So far, the group has been making steady progress according to the Gantt chart. There are minor changes in terms of contribution of work, we got to start some topics earlier than expected but finished later than expected, and they are mainly due to the complexity of data for clustering algorithms and visualizations. The modified and original Gantt charts are shown below. All members contributed evenly.



TASK	TOPIC	RESPONSIBLE MEMBER	PROGRESS	START	END
1	Data Cleaning and Preping	Fei & Sufian	50%	2/27/23	3/4/23
2	Experiment and Workflow Design	Yuxin, Rutuja & Alessio	0%	2/27/23	3/4/23
3	Ranking Algorithm on Ingredients	Yuxin, Rutuja & Alessio	0%	3/5/23	3/15/23
4	Classification Algorithm on Healthy/Not	Fei & Sufian	0%	3/16/23	3/25/23
5	Progress Report	Alessio & Sufian	0%	3/26/23	3/30/23
6	Interactive Visualizations	Alessio & Yuxin	0%	4/1/23	4/10/23
7	Final Result Validations and Analysis	Fei & Rutuja & Sufian	0%	4/1/23	4/10/23
8	Final Report	All 5 members	0%	4/10/23	4/20/23

Figure 1: Original Gantt Chart



TASK	TOPIC	RESPONSIBLE MEMBER	PROGRESS	START	END
1	Data Cleaning and Preping	Fei, Sufian	100%	2/27/23	3/4/23
2	Experiment and Workflow Design	Yuxin, Rutuja	100%	2/27/23	3/4/23
3	Ranking Algorithm on Ingredients	Fei, Sufian	100%	3/5/23	3/20/23
4	Progress Report	All 5 members	100%	3/26/23	3/30/23
6	Data Visualization infrastructure (network graph, heatmap)	Alessio	100%	3/5/23	3/30/23
5	Classification Algorithm on Healthy/Not	Yuxin, Rutuja	50%	3/20/23	4/7/23
7	Interactive Visualizations	Alessio, Fei, Sufian	80%	3/5/23	4/10/23
8	Final Result Validations and Analysis	All 5 members	0%	4/7/23	4/13/23
9	Final Report	All 5 members	0%	4/14/23	4/20/23

Figure 2: Modified Gantt Chart

Conclusions & Discussion. The biggest challenge will be to evaluate the results of the clustering algorithm. The important thing to note is that the nature of unsupervised learning paves the way for a more generalized result which provides creative liberties but also limits a source of comparison and verification for accuracy. Our approach for this is to use highly useful websites and resources [19] [20] in order to compare and contrast the results of our algorithm with that of those on the web. For example, our aim is having our algorithm rank items like chocolate chip cookies as lower on the health spectrum than raw carrots due to the limited processing and calories. Since this fact is also verified by our supplemental websites, our method of evaluation will be to pull results from this source and compare it

to our ranking. Using a quick numerical analysis that matches the order of our items to the order of their items, we'll be able to discover the success of our algorithm. In addition, the results of the classification currently are weighing heavily on the caloric aspect of the data based on the analysis. While 5 clusters have been developed, we still need to perform the verification step with the resource we've found to somewhat validate the clustering model. In addition, another approach we may explore is the concept of self-labeling the data based on scientific research found. For example, an item will be labeled in the dataset as "healthy" or "unhealthy" based on a range of values for each of the categories (ie. An item is considered as containing too much fat due to the high levels of saturated fat) and using this categorization, we will thus use this to compare the results of the clustering algorithm. While this approach is an additional step and potentially biased based on the research we conduct, it is an approach we will explore for the sake of validating our model.

Literature Survey Scientific approaches such as the Nutrient Profiling System (NPS) involve the careful examination of every single ingredient and its nutritional value [12]. More visual efforts addressing allergies and dietary concerns have been tackled by CNN, through AlexNet, using specific image datasets for training and classification [17]. Kazama [8] combines a barycentric Newton diagram algorithm and Word2vec model to substitute food ingredients with alternative belonging to another region and/or culinary culture. By leveraging the different types of food additives and their purposes discussed by Spencer [16], we will try to integrate our ranking-by-popularity algorithm, with a further classification between healthy vs unhealthy for each ingredient discovered. Similarly, Moubarac [12] explains how several food classification systems have been produced, studied, and evaluated. Despite a seemingly comprehensive approach, these studies fail to address the evolution of industrial food processing over the years, a gap that we will try to fill in our study. For visualization, Dunford [5] unveils FoodSwitch, an app that provides easy-to-understand nutritional information to Australian consumers to raise awareness and lead to healthier food choices. The visualization could be improved with map ingredients and products. Chatterjee et al. [1] demonstrate how the Formal Concept Analysis (FCA) technique can be used in food ingredient analysis to show relationship between data category and attributes. Similarly, we will obtain food ingredients from a neural-network of grocery-store-specific items. By increasing the amount of information conveyed, and at the same time lowering the intellectual foundations needed to interpret such information, Park [15] believes that data visualization proves to be the better medium for large scale communication. We will implement these concepts throughout, with the final goal of delivering accurate at-a-glance information. The model KitcheNette, developed by Donghyeon [2], helps predict food ingredient pairing scores and recommends optimal ingredient pairings using Siamese neural networks. The pairing relationship could be better understood by our ingredient-network graph. Kim and Chuan [9] propose a knowledge-based hybrid decision model using supervised learning neural networks for nutrition management. The model is a food recommendation system that helps consumers make healthier dietary decisions by collecting food preference data. Kirk [10] sustains that machine learning algorithms have the potential to supplement traditional approaches to nutrition research, a claim that could be supported by statistical results. Many of the proposed validation algorithms could be useful in our results with high-dimensional data. Drewnowski [3] explains the Food Compass Score (FCS): a nutrient profiling system. This study postulates that food products that carry more nutritional value compared to calories, are defined as nutrient dense. We can adopt this methodology to provide meta data for our results. According to Juul et al. [7], families that tend to have minimal ingredients in their items in the grocery cart tend to have healthier outcomes. While this isn't a direct implication, it can suggest that an excessive amount of ingredients may equate to highly processed foods and therefore unhealthy when compared to minimal ingredient foods such as vegetables and fruits. A paper by Drewnowski [4] to help consumers identify healthier food at an affordable price. The NRF index, used with a database of food prices, can serve as a health guide for end users. While Food Compass, a brand-

new nutrient profiling system developed by O' Hearn [14] incorporates 54 food attributes across nine different domains to generate a score that ranks individual food items on a scale of 1-to-100 unhealthy (low) vs healthy (high) values. A probabilistic sample on the Brazilian territory, helped Fellegger Garzillo [6] determine the relationship between ultra-processes food products and carbon food print. This information will likely raise awareness and perhaps alter the dietary habits of consumers. A group of researchers [11] have analyzed carbon footprint in relation to dietary guidelines over a world spread study region. This will shine light on which countries contribute more to the problem and how food consumption is directly associated with it.

References

1 Chatterjee, U. et al. (2016). Formalizing Food Ingredients for Data Analysis and Knowledge Organization. *Collnet Journal of Scientometrics and Information Management.*

2 Donghyeon, P. et al. (2019) KitcheNette: Predicting and Recommending Food Ingredient Pairings using Siamese Neural Networks. Retrieved from: https://arxiv.org/pdf/1905.07261.pdf

3 Drewnowski A, Fulgoni. VL. (2014). Nutrient density: principles and evaluation tools. *Am J Clin Nutr*. Retrieved from: https://pubmed.ncbi.nlm.nih.gov/24646818/

4 Drewnowski A. (2010). The Nutrient Rich Foods Index helps to identify healthy, affordable foods. *The American Journal of Clinical Nutrition*, Volume 91

5 Dunford, et al. (2014). FoodSwitch: A Mobile Phone App to Enable Consumers to Make Healthier Food Choices and Crowdsourcing of National Food Composition Data. *JMIR Mhealth Uhealth*. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4147708/

6 Fellegger Garzillo, J.M. et al. (2022). Ultra-processed food intake and diet carbon and water footprints: a national study in Brazil. *Rev Saude Publica.* Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8859933/

7 Juul, F. et al. (2019). Processing level and diet quality of the US grocery cart: is there an association? Cambridge University Press

8 Kazama, M. (2018). A Neural Network System for Transformation of Regional Cuisine Style. *Sec. Big Data Networks.* Retrieved from: https://www.frontiersin.org/articles/10.3389/fict.2018.00014

9 Kim, J.C., Chuang, K. (2020). Knowledge-based hybrid decision model using neural network for nutrition management. *Inf Technol Manag 21*. Retrieved from: https://link.springer.com/article/10.1007/s10799-019-00300-5

10 Kirk, D. et al. (2022). Machine Learning in Nutrition Research. Advances in Nutrition, Vol. 13.

11 Kovak, B. et al. (2021). The carbon footprint of dietary guidelines around the world: a seven country modeling study. *Nutrition Journal Vol. 20.*

12 Moubarac, J.C. et al. (2014). Food Classification Systems Based on Food Processing: Significance and Implications for Policies and Actions: A Systematic Literature Review and Assessment. *Current Obesity Reports* Vol. 3

13 Mozaffarian, D. et al., (2021). *Food Compass is a nutrient profiling system using expanded characteristics for assessing healthfulness of foods.* Nature Food.

14 O' Hearn, M. (2002). Validation of Food Compass with a healthy diet, cardiometabolic health, and mortality among U.S. adults, 1999–2018. *Nat Commun 13*. Retrieved from: https://doi.org/10.1038/s41467-022-34195-8

15 Park, S. et al. (2022). Impact of data visualization on decision-making and its implications for public health practice: a systematic literature review. *Informatics for Health and Social Care*. Retrieved from: https://www.tandfonline.com/doi/abs/10.1080/17538157.2021.1982949?journalCode=imif20

16 Spencer, M. (1974). Food Additives. *Postgraduate Medical Journal.* Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2495648/pdf/postmedj00310-0028.pdf

17 Qiaomei, Z. et al. (2019). Comprehensive Reviews in Food Science and Food Safety. Vol. 18

18 Wolff, K.E. (1996). A first course in Formal Concept Analysis: How to Understand Line Diagrams. Retrieved from: https://sites.tufts.edu/ancientbirds/files/2018/06/a-first-course-in-formal-concept-analysis.pdf

ǂ1 Datafiniti (2017). Food Ingredient Lists. Retrieved from: https://www.kaggle.com/datasets/datafiniti/food-ingredient-lists?resource=download&select=ingredients+v1.csv
ǂ2 NHS Food Facts. https://www.nhs.uk/healthier-families/food-facts/
ǂ3 Open Food Facts. https://world.openfoodfacts.org/.