CSE 6242 Fall 2021

THE MOODY FOODIES

Team # 58

Final Report

Ibukun Ajayi, Monica Cheng, Justin Huang, Yang Luo, Justin Nand

ABSTRACT

Previous methods of generating meal alternative recommendations relied on the analysis of food ingredients using techniques such as collaborative filtering based on ingredient co-occurrence. In this work we present a nearest neighbor search based on neural embeddings generated from the flavor molecules present in a meal recipe while optimizing for nutritional constraints. Additionally, the relationships between flavor vectors are explored to elucidate the models understanding of “flavor”. The trained “flavor space” embeddings show a top-25 identity accuracy of 83% using the doc2vec PV-DBOW model, while qualitatively passing tests for food alternatives with similar flavor profiles – implying the model’s working understanding of “flavor”.

INTRODUCTION

When considering the satiating effects of food, flavor is often secondary to the macronutrient profile of a meal where alternatives with more desirable nutritional profiles are selected with little consideration to taste. We propose making flavor a first-class citizen, then amongst similar tasting foods search for the best nutritional choice. Our intuition suggests that a more pleasant eating experience will lead to increased overall happiness, healthier choices, and fewer dieting “cheat meals”.

Our project aims to create a recommendation system based on perceived “taste” profile while surfacing a nutritionally optimal alternative – retaining the psychologically satiating effects of “flavor”.

PROBLEM DEFINITION

60% of American adults have one or more diet-related chronic diseases according to the USDA [1]. Dietary recommendations from experts aim to reduce disease and improve health of the general population, however, translating dietary guidelines into practical and achievable healthy food choices remains a challenge. Food consumption patterns, prices, cultural-specific diets, and health condition have all been identified as potential barriers for adopting new eating patterns [2]. Much research has been dedicated to designing new food intake patterns which optimize the nutrient content while maintaining original or culture-specific consumption patterns as much as possible [3] [4]. Mathematical diet optimization has been applied but may not be accessible. Studies have combined nutrition optimization with recipe recommendation to provide more accessible health food alternatives. However, some of the algorithms employed by these models rely on manual selection of key features where the importance of a feature depends solely on a single human curator’s perception of a meal alternative [5] [6].

RELATED WORK / LITERATURE SURVEY

The 1M+ recipe set is the largest publicly available corpus of recipes [7]. A subset of that data was annotated with nutritional information. Work from Garg et al. created Flavor DB [8], a database of molecules that contribute flavors and odors to the culinary ”flavor” experience. Using these two datasets Park et al. constructed Flavor Graph, representing the co-occurrence of ingredients in the recipe 1M+ dataset and their relationship to flavor molecules [9]. Work from Simas et al. [10] explores the novel concept of food-bridging as compared to the traditional method of food pairing, where recipes are developed by mixing ingredients with compatible chemical compounds [11]. The Food-bridging hypothesis assumes that if two ingredients do not share a strong molecular or empirical affinity, they may become affine through a chain of pairwise affinities. Simas et al.’s hypothesis sets the ground work for our own hypothesis of classically-condition psychological perceptions of ”flavor”.

Natural Language Processing (NLP) provides the computational foundations of our method, enabling data integration and analysis of the recipe and flavor datasets. This work explores application of the Continuous Bag-of-Words (CBOW) and Skip-gram models adapted in doc2vec of Le et al. Doc2vec removes non-linear

layers, reducing computational complexity during training compared to more expensive Recurrent Neural Net (RNN) language models and transformer based methods [12]. Other models offer increasing levels of robustness, at the cost of training and inference time. Ingredient substitution has been demonstrated using ingredient embedded vectors to identify similar contexts [15]. Teng, et al was able to deduce functionally equivalent ingredients using a complement network and a substitution network on the All Recipes dataset [16].

METHODS

Our intuition leverages advances in natural language processing's ability to surface high level concepts that generalize relationships between words in a document corpus for understanding the psychological experience of "flavor" using the molecular basis of flavor. Similar foods can be easily identified by their shared flavor molecules, but we believe that satiating meal alternatives exist that may not share the same molecular flavors but due to co-occurrence with "out of profile" molecules can simulate a classically-conditioned psychological effect using flavor molecules that are frequently experienced with a canonical flavor molecule – effectively capturing the abstract concept of "flavor".

DATA GATHERING & PREPARATION

This work used a subset of the Recipe 1M+ dataset [7] which includes 50,000 recipes annotated with nutritional data. Molecular flavor profiles of known food ingredients from FlavorDB's web interface [8] were downloaded with a developed python script. The recipe documents were joined to the molecular flavor data via a recipe-to-ingredient-to-molecular-flavor entity relationship, using ingredient names as lookup keys in FlavorDB.

Since the ingredient list in the Recipe 1M+ dataset is not tokenized in a form consistent with the ingredient names in FlavorDB, we then normalized, string-matched, and tokenized ingredients between the two datasets to effectively join them. Our approach tested variants of string matching methods based off of Levenshtein distance [17] [18]. We benchmarked the matching algorithms on their accuracy and speed, choosing the technique that provided the highest quality result.

Once the two datasets were joined, they were formatted to follow an expected schema, with one document per recipe. The joined molecular data was then transformed such that the list of flavor molecules from each component ingredient was merged into a single set where each molecule's presence was uniquely recorded by its PubChem ID. The consolidated recipe flavor molecule lists serve as inputs into doc2vec.

EMBEDDINGS

Further processing of the input was handled by the Gensim library, which takes all the input molecule lists and composes a vocabulary that one-hot encodes the molecules [19]. To model the recipe-to-molecule relationship, we apply the doc2vec extension of the word2vec algorithm. Doc2vec effectively captures the compositional nature of the recipe-to-flavor-molecule relationship by appending a recipe ID vector to the training input [20]. Doc2vec provides two modeling options, distributed-memory, and distributed-bag-of-words. Both models were trained using a window size of 772, based on the max length of all molecule lists, a vector size of 300, for 40 epochs. Once the models were trained embeddings for every recipe and ingredient were computed and stored for nearest neighbor search.

NEAREST NEIGHBOR SEARCH

Using the computed recipe embeddings, we performed nearest neighbor search using Gensim's built in "most_similar" API, which internally implements a cosine similarity-based vector search. The search neighborhood was limited to the top 10 closest recipes by highest similarity. The success of similarity search as a proxy for meal recommendation was evaluated by taking each food items molecular list and

re-inferring an embedding, then searching for its nearest neighbors, taking note of the top-k depth need to locate an identity match to the queried food item. The findings were plotted as an ROC curve.

Dimensional Reduction & Clustering

Exploratory analysis of the "flavor space" was done by applying "Uniform Manifold Approximation and Projection" (UMAP) to reduce the dimensionality of the space to three dimensions [21]. The resulting points were then clustered using the "Density-Based Spatial Clustering of Applications with Noise" (DBSCAN) algorithm [22]. The clusters were then examined for emergent relationships based on annotation term enrichment and their unique flavor vs other clusters in the "flavor space". Unique defining flavors/notes for each cluster were surfaced by normalizing the flavor annotation terms by applying a Term Frequency – Inverse Document Frequency (TF-IDF) transformation.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Where a word/term/flavor i in cluster j is scored as the product of its term frequency (tf) and the log of the total number of clusters (N) divided by the number of clusters containing term i (df). Term frequency is calculated as the count of the term divided by the total number of terms in a cluster.

Optimization & Ranking

The Recipe1M+ dataset contains the information of the total nutrition values per 100g for each recipe. The nutrients included in the data are total calories, fat, protein, salt, sugar, and saturated fat. To optimize our output recommendations, we followed the daily nutrition consumption recommendation (Daily Value, DV) from the FDA as a standard value and then calculated a redistributed profile based on category weights. Using FDA DVs as the recommended amounts of nutrients to consume or not to exceed each day, we normalized a single serving of an individual food based on its weighted contribution to one's daily diet as the benchmark value.

$$\%\mathrm{DV} = \frac{nutrition\ value\ per\ 100g\ of\ the\ food}{daily\ nutrition\ recommend\ value\ (DV)}$$

Using the %DV for each nutrient, we can then calculate a single healthiness score for each recipe with the following formula:

$$\mathrm{H} = c_1\%DV_{Fat} + c_2\%DV_{protien} + c_3\%DV_{salt} + c_4\%DV_{saturate} + c_5\%DV_{sugar}$$

where c1, c2, c3, c4, c5 are constants that a user can pick depending on their dietary needs. We established a default setting, with c3, c4, c5 < 0 and c2 > 0 under the guidance that we want to minimize sodium, saturate fat and sugar content while maximizing protein content. Additionally, we provided modes for "Ketogenic", and "Low Fat" diets.

Experiments and Evaluations

Ingredient Tokenization

Different methods of string matching were evaluated based on variants of Levenshtein distance in the fuzzywuzzy library. To evaluate the accuracy, a validation set of 100 recipes from 1M+ was selected and the 626 ingredients within the 100 recipes were manually matched to their most representative entries in FlavorDB. Once these true labels were determined, each string match algorithm could be evaluated by scoring percent accuracy of the prediction on the validation set.

Comparison of different ratios for scoring were evaluated including whole string, partial string, token sort, token set, and a weighted ratio that combines all 4 ratios. In addition, several truncation points of input ingredient strings separated by commas were tested. Accuracy was also evaluated on top k returned matches as another metric of algorithm's performance (Figure 1). Weighted ratio with string truncation of 2 performed with 80% accuracy and was used to join our dataset.

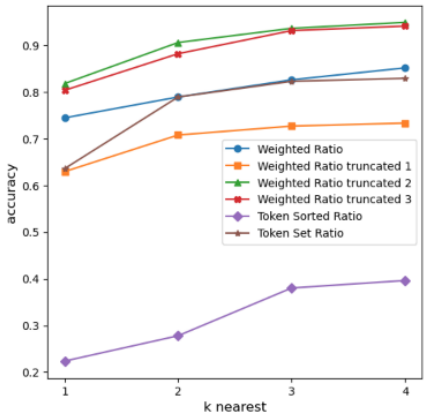

FIGURE 1 - STRING MATCH ACCURACY

Mismatches were analyzed with the aid of confusion matrix to determine sources of mismatch error (Figure 2). Most common source of mismatch was the inability to distinguish between close matches such as matching 'wheat flour' to 'wheat' instead of 'flour'. Other sources included different spellings 'catsup' instead of 'ketchup'. Some ingredients did not have existing matches in FlavorDB such as 'fish sauce'.

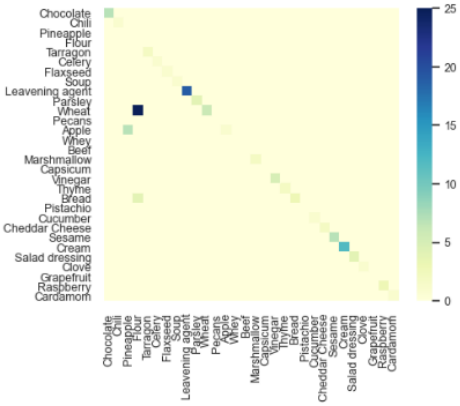

FIGURE 2 - CONFUSION MATRIX

DATA QUALITY

Among the 51,235 recipes, 52 records contain only 1 nutrient value, 35 with only 2 nutrient values. These records which contain limited nutrition data are either none-food recipes that somehow appear on the recipe website (such as Anti-Itching Cream) or recipes missing ingredient amounts.

To further estimate the error size of the data, we randomly sampled 382 records (in order to reflect the actual dataset with 95% confidence level) from the 51,235 recipes and manually investigated the content. The investigation shows that, unfortunately, the value of the amount per ingredient is handled incorrectly for all the recipes originally extracted from www.food.com. When an ingredient has an amount in fraction (1/2, etc.), the number was pulled missing the "/". That is 1/2 will become 12. This error can inflate the nutrient value for such ingredient and thus bias the calculation for total nutrient contain for such recipe. Since www.food.com contributes about 55% of the total recipes in the 1M+ data subset an alternative source for the nutrition data must be considered if we want to extend this project in the future.

EMBEDDING MODELS

Both variants, PV-DM and PV-DBOW, of the doc2vec model were trained using a 25 core 100 GB compute server. PV-DM took **~55 CPU hours** to train, while PV-DBOW took only **6 CPU minutes**.

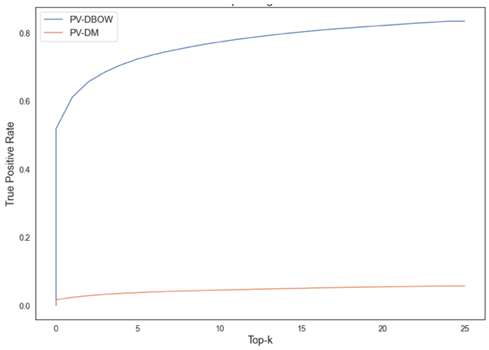

FIGURE 3 - ROC PV-DM VS PV-DBOW

An identity search was carried out by increasing the top-k results incrementally up to 25 as described in the methods. The true positive "identity" hit rate was plotted as a ROC curve (Figure 3), and showed PV-DBOW to be far superior to PV-DM.

Clustering analysis showed PV-DM to produce **2 clusters** (Figure 4)., while PV-DBOW had **437 clusters** (Figure 5). Examination of the dimensionally reduced embeddings shows the PV-DBOW model to produce more distinct clustering with

better proximal recommendations while PV-DM learned a continuous manifold. Based on these results the PV-DBOW model was selected for computing the flavor space used in downstream analysis.

Nutritional Optimization

The 1M+ recipe data includes the UK FSA traffic light label to indicate the healthiness level for each recipe. The traffic light label can be converted into a single score to indicate the overall healthiness of a recipe. However, this discrete scoring method may not be sufficient to distinguish smaller differences of nutrition value nor to provide the flexibility for differentiating different types of dietary preference such as Keto diet. However, the current optimizer is also more vulnerable to extreme values.

"Computational Cooking"

To determine if the model understands the perceptual concept of "flavor", we performed three qualitative "cooking" tasks.

"Tastes like chicken"

The first of which searches for foods that "taste like chicken" by querying the chicken ingredient flavor vector for its nearest neighbors and then filtering out any recipes that explicitly included chicken in its ingredients list. The results suggested many turkey-based recipes with top similarity.

What is "meat" flavor?

The second task attempted to determine what molecules capture "meat" flavor by intersecting the molecule lists of Chicken, Turkey, Pork, Beef, and Lamb. This identified 3 molecules. A flavor vector was inferred from the 3 molecules, and its nearest neighbors found. The results included many types of meat tasting suggestions: sausage, ham, buttermilk, whitefish, walrus, and a macaroni.

Flavor math

The third task used the flavor vectors for peanut, honey, potato, and the meat vector found previously to check if vector math operations would yield intuitive recommendations. The query vector was computed as follows:

Query = Peanut + Honey + Meat Flavor – Potato

The query's nearest neighbors resulted in many sauces, but one result in particular stood out as validation of the flavor math intuition: "Honey-Glazed Pecans" with a rank position 4.

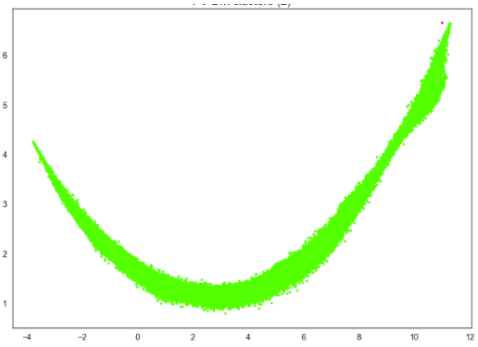

FIGURE 4 - CLUSTER PV-DM

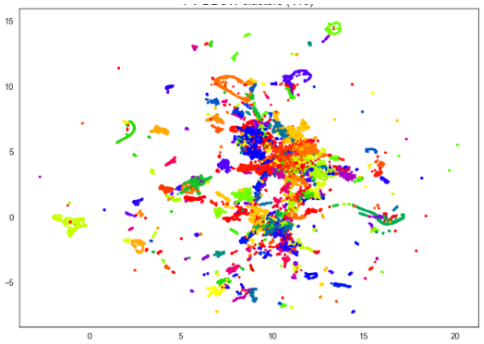

FIGURE 5 - CLUSTER PV-DBOW

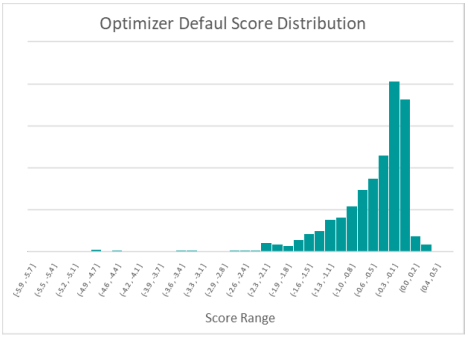

FIGURE 7 - OPTIMIZER SCORE DISTRIBUTION

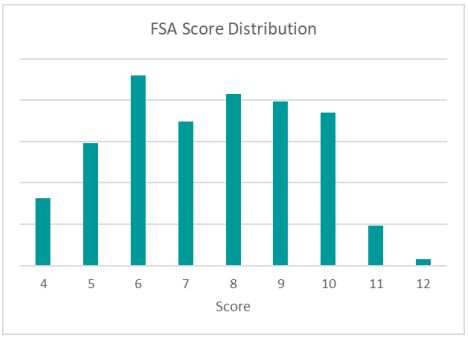

FIGURE 6 - FSA SCORE DISTRIBUTION

VISUALIZATION & USER EXPERIENCE

Our work provides two graphical user interfaces (GUI) for interactively exploring the flavor space.

The first is a cluster browser that allows a user to click on one of the clusters in its left panel and view a word cloud visualization of the cluster's taste profile.

Additionally, a TF-IDF normalized word cloud is displayed alongside the full taste profile to show any cluster specific distinctive flavors. This interface is very useful for confirming the embedding model's ability to distinguish flavor themes from one another and identify large cluster that have relatively few flavor molecules as dominate distinctive flavors.

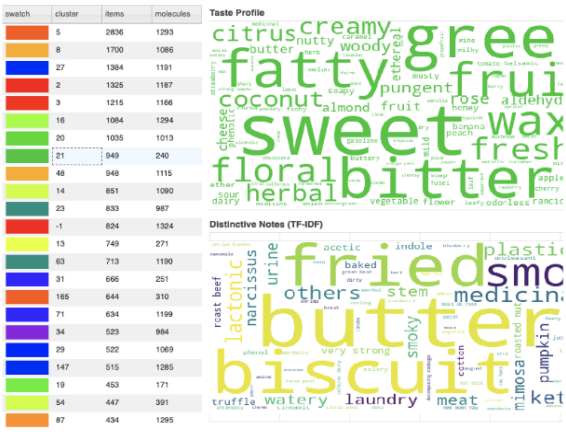

FIGURE 8 – CLUSTER BROWSER

The second GUI displays an interactive plot of flavor space, allowing the user to visually inspect clusters, and verify the non-uniformity of molecular features in the space. Additionally, a user can select a food item to query against flavor space returning its nearest neighbors with similar flavors. The Nutritional priority field applies our optimization method to the recommended results and selects the best option for a given diet profile.

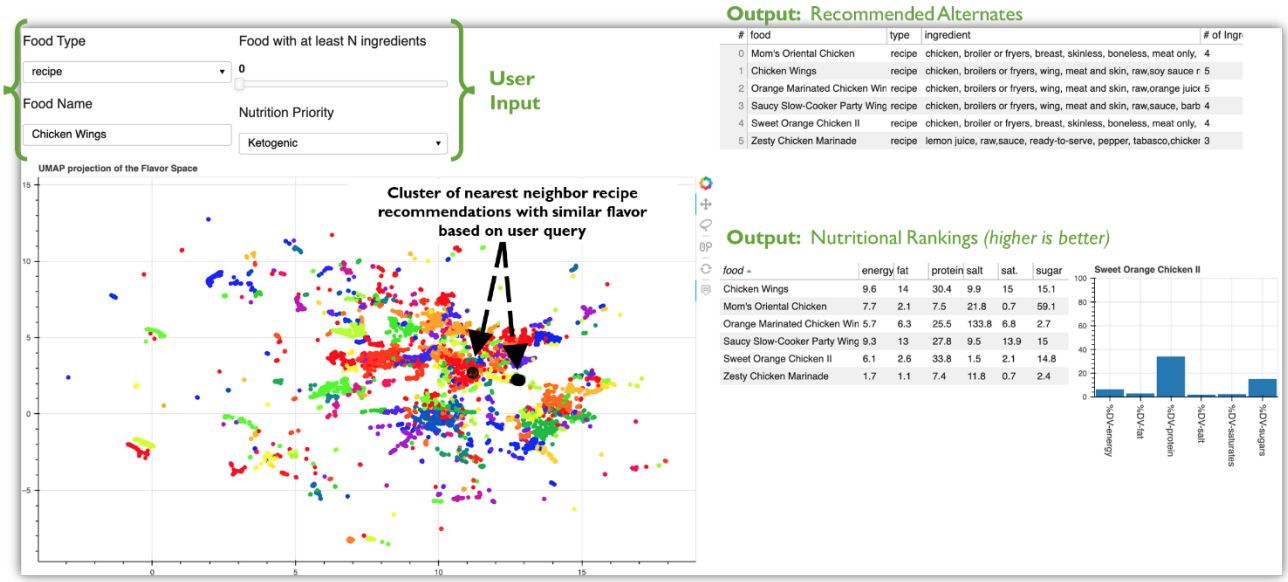

FIGURE 9 – FLAVOR SPACE EXPLORER & MEAL ALTERNATIVE RECCOMENDER

CONCLUSIONS AND DISCUSSIONS

While there were many sources of error and noise contributing to spurious results, we found the model robust enough to still effectively capture the conceptual idea of "flavor" in terms of the molecular flavor composition of food. Most impressive is the model's ability to provide flavor recommendations in cases where there are no shared ingredients between the query and its suggested alternatives – exemplified by the "tastes like chicken" test. We see this work as means to achieve healthy diet alternatives without compromising on the experience of eating. Further work might investigate the model's suggestion of psychological flavor analogs as alluded to in the introduction.

CONTRIBUTIONS

All members contributed equally to this work.

REFERENCES

[1] "Dietary Guidelines for Americans, 2020-2025," U.S. Department of Agriculture.

[2] Lee M Ashton, Melinda J Hutchesson, Megan E Rollo, Philip J Morgan, Debbe I Thompson, Clare E Collins, "Young adult males' motivators and perceived barriers towards eating healthily and being active: a qualitative study," *International Journal of Behavioral Nutrition and Physical Activity,* p. 93, 2015.

[3] Hitomi Okubo, Satoshi Sasaki, Kentaro Murakami, Tetsuji Yokoyama, Naoko Hirota, Akiko Notsu, Mitsuru Fukui, Chigusa Date, "Designing optimal food intake patterns to achieve nutritional goals for Japanese adults through the use of linear programming optimization models," *Nutrition Journal ,* p. 57, 2015.

[4] Gabriel Masset, Pablo Monsivais, Matthieu Maillot, Nicole Darmon, Adam Drewnowski, "Diet Optimization Methods Can Help Translate Dietary Guidelines into a Cancer Prevention Food Plan," *The Journal of Nutrition,* vol. 139, no. 8, p. 1541–1548, 2009.

[5] Youri van Pinxteren, Gijs Geleijnse, Paul Kamsteeg, "Deriving a recipe similarity measure for recommending healthful meals," in *Proceedings of the 2011 International Conference on Intelligent User Interfaces*, Palo Alto, CA, USA, 2011.

[6] Shuyang Li, Julian McAuley, "Recipes for Success: Data Science in the Home Kitchen," Harvard Data Science Review, 2020.

[7] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, Antonio Torralba, "Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 43, no. 1, pp. 187-203, 2019.

[8] Neelansh Garg, Apuroop Sethupathy, Rudraksh Tuwani, Rakhi NK, Shubham Dokania, Arvind Iyer, Ayushi Gupta, Shubhra Agrawal, Navjot Singh, Shubham Shukla, Kriti Kathuria, Rahul Badhwar, Rakesh Kanji, Anupam Jain, Avneet Kaur, Rashmi Nagpal, Ganesh Bagler, "FlavorDB: a database of flavor molecules," *Nucleic acids research,* vol. 46, no. D1, pp. D1210-D1216, 2018.

[9] Donghyeon Park, Keonwoo Kim, Seoyoon Kim, Michael Spranger, Jaewoo Kang, "FlavorGraph: a large-scale food-chemical graph for generating food representations and recommending food pairings," *Scientific Reports,* vol. 11, no. 1, pp. 1-13, 2021.

[10] Tiago Simas, Michal Ficek, Albert Diaz-Guilera, Pere Obrador, Pablo R. Rodriguez, "Food-Bridging: A New Network Construction to Unveil the Principles of cooking," Frontiers in ICT, 2017.

[11] Ayana Dagan-Wiener, Ido Nissim, Natalie Ben Abu, Gigliola Borgonovo, Angela Bassoli, Masha Y. Niv, "Bitter or not? BitterPredict, a tool for predicting taste from chemical structure," *Scientific Reports,* vol. 7, no. 1, pp. 1-13, 2017.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv, 2013.

[13] Diya Li, Mohammed J. Zaki, "RECIPTOR: An Effective Pretrained Model for Recipe Representation Learning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention Is All You Need," Advances in neural information processing systems, 2017.

[15] Dennis Lawo, Lukas Böhm, Gunnar Stevens, "Veganaizer: AI-assisted Ingredient Substitution," 2020.

[16] Chun-Yuen Teng, Yu-Ru Lin, Lada A. Adamic, "Recipe recommendation using ingredient networks," in *Proceedings of the 4th Annual ACM Web Science Conference*, 2012.

[17] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady,* vol. 10, no. 8, pp. 707-710, 1966.

[18] SeatGeek Inc, "Fuzzywuzzy: Fuzzy String Matching in Python," 2014. [Online]. Available: https://github.com/seatgeek/fuzzywuzzy.

[19] Rehurek, Radim and Sojka, Petr, "Gensim--python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic,* vol. 3, no. 2, 2011.

[20] Le, Quoc and Mikolov, Tomas, "Distributed Representations of Sentences and Documents," *International Conference on International Conference on Machine Learning,* vol. 32, p. II–1188–II–1196, 2014.

[21] McInnes, Leland and Healy, John and Saul, Nathaniel and Grossberger, Lukas, "UMAP: Uniform Manifold Approximation and Projection," *The Journal of Open Source Software,* vol. 3, no. 29, 2018.

[22] Ester, Martin and Kriegel, Hans-Peter and Sander, J"org and Xu, Xiaowei and others, "A density-based algorithm for discovering clusters in large spatial databases with noise.," *kdd,* vol. 96, no. 34, pp. 226-231, 1996.