Supporting Plant-based Diets with Ingredient2Vec

Dennis Lawo
Information Systems
University of Siegen
Siegen, Germany
Email: dennis.lawo@uni-siegen.de

Lukas Böhm
Information Systems
University of Siegen
Siegen, Germany
Email: lukas.boehm@uni-siegen.de

Margarita Esau
Information Systems
University of Siegen
Siegen, Germany
Email: margarita.esau@uni-siegen.de

***Abstract*—Plant-based diets have taken on an entirely new significance as the ecological consequences of diet choice have become more apparent; it is now acknowledged that dietary choices have significant consequences for sustainability. However, plant-based cooking and the 'veganization' of recipes is something newcomers to these new cuisines struggle with. Attempting to support sustainable food choices and the learning of plant-based cooking, we propose a Word2Vec-based approach for AI-assisted recipe 'veganization'. In this paper, we describe the fundamental thoughts behind the algorithm and explore challenges and opportunities for future research in a short case study.**

I. Introduction

Plant-based diets, with respect to climate change, have taken on an entirely new significance as the ecological consequences of diet choice have become more apparent. One challenge of supporting people in adopting plant-based diets lays in the acquisition of the relevant knowledge about substitutes and how to incorporate them into well-known and beloved recipes [1]. While a lot of recipes and food-related information is nowadays acquired from the internet, the food computing community missed the opportunity to research ingredient substitution and therefore support plant-based dieting [2]. However, word embeddings[3] are promising new directions to examine for solving such substitution tasks[4].

Building upon the usage of word embeddings, this paper explores first opportunities and challenges from the implementation of an Ingredient2Vec substitution algorithm. The fundamental task the model aims to solve is finding the most likely substitute s given the ingredient i, whereby $s \in I_v$ and $s \in V_o$ with I_v as the vocabulary containing vegan ingredients and I_o as the vocabulary containing omnivorous ingredients. In other words, the task is solving: 'Meat' is for 'Omnivorous Cuisine' as *'Tofu'* is for 'Vegan Cuisine'.

II. Related Work

The search for substitutes between different cuisine styles is an underrepresented and new area of research in food computing [2]. While some work used static taxonomies of substitutes [5], current research uses vector embeddings generated by non-negative matrix factorization [6] or singular value decomposition [7]. While they were generally able to find similar ingredients, their work did not focus on a 'goal-directed' substitution of ingredients. Here the usage of Word2Vec to solve analogies, similar to the above example, showed first promising results in the food domain [4]. Using a regional cuisine feature Kazama et al. [4] were able to translate between regional cuisines, e.g. 'Mirin' is for 'Japanese' as *'Calvados'* is for 'French'. However, an evaluation of the capabilities and barriers of this approach as well as an exploration of the application scenarios is still missing.

III. Ingredient2Vec Approach

A. Data Retrieval & Preprocessing

The recipes were scraped from two popular German recipe websites[1][2] and processed through a pipeline that extracted the ingredient table and the categorization as vegan and omnivorous from the raw page source. Given the ingredients and instructions of 452,277 recipes, we started cleaning the data based on heuristic and string matching approaches [8]. Based on this clean representation of recipes, duplicates, ingredients with less than 15 occurrences and recipes smaller or equal than 4 ingredients have been removed, resulting in 384,181 recipes with 4,116 distinct ingredients.

B. Ingredient2Vec

We adapted the original Word2Vec algorithm [3] in line with the research of Kazama et al.[4]. In particular, we trained the neural network (see Figure 1) with ingredients w_i and categories indicating 'vegan' or 'omnivorous' c_r, such that the probability of occurrence is given a an output. Therefore we used the objective function:

$$\sum_{r \in R} \sum_{w_i \in r} (logP(w_i|c_r) + logP(c_r|w_i) + \sum_{j \neq i} logP(w_j|w_i)) \quad (1)$$

with the probability defined as:

$$P(b|a) = \frac{exp(v_a^T v_b')}{\sum_{c \in W} exp(v_a^T v_c')} \quad (2)$$

with $a, b, c \in Ingredients \cup Categories$ and $v_a, v_a' \in RK$ as the k-dimensional vector representations of the ingredient/country. After training, the hidden layer vector representation of the ingredients and categories are then used to solve analogies in the form of 'Meat' is for 'Omnivorous Cuisine' as *'Tofu'* is for 'Vegan Cuisine', by calculating $v_{meat} - v_{omnivorous} + v_{vegan} = v_{predict}$ and finding the matching ingredient from $i \in Ingredients$ where

[1] kochbar.de/

[2] chefkoch.de/ (vegan recipes only)

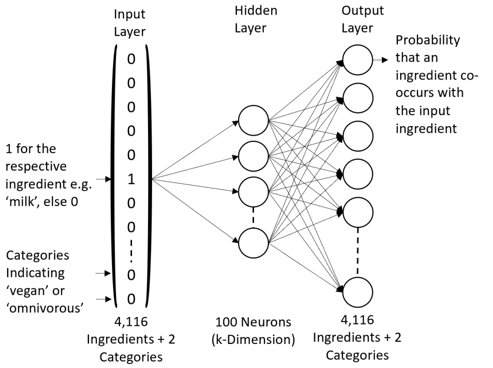
Input Layer
Hidden Layer
Output Layer
Probability that an ingredient co-occurs with the input ingredient
1 for the respective ingredient e.g. 'milk', else 0
Categories Indicating 'vegan' or 'omnivorous'
0
0
0
0
0
1
0
0
0
0
4,116 Ingredients + 2 Categories
100 Neurons (k-Dimension)
4,116 Ingredients + 2 Categories

Fig. 1. Ingredient2Vec Neural Network

$max(cosineSimilarity(v_i, v_{predict}))$. In line with Kazama et al. [4] we have chosen $k = 100$. The implementation was conducted using Python and Gensim[3].

IV. Evaluation

In the following, we want to provide some examples showing the capabilities of the approach and how it can be used in recipe websites and applications to support people with sustainable food choices. However, since there is a lot of space for improvements we also discuss arising challenges for future work.

A. Performance for Selected Results

In the following some selected examples of predictions made by the model are presented:

- 'Sausage' is for 'Omnivorous' as *'Smoked Tofu'* is for 'Vegan'
- 'Milk' is for 'Omnivorous' as *'Soy Milk'* is for 'Vegan'
- 'Cheese' is for 'Omnivorous' as *'Vegan Cheese'* is for 'Vegan'
- 'Cream' is for 'Omnivorous' as *'Soy Cream'* is for 'Vegan'
- 'Honey' is for 'Omnivorous' as *'Agave Syrup'* is for 'Vegan'

On the first sight those results look very reasonable. What is interesting is that the model does not just suggest one ingredient but a ranked list of various substitutes. For example within the top-10 recommendations for 'cream' we also find 'rice cream' and 'oat cream'. However, there are multiple different ingredients where the model is yet not performing well, e.g. for 'pork' the top-recommendation is 'tandoori masala' with 'textured vegetable protein' at rank 2' or for 'egg' none of the top-5 recommendations seem reasonable.

B. Beyond Probability

A further challenge arises from the chemical characteristics of some recommendations made. For the various 'meat' substitutes it is rather a matter of personal preference if one wants to fry 'tofu', 'textured vegetable protein steak' or 'tempeh'. However, for 'egg' the chemical characteristics of the substitute are more important and context-dependent. While for baking the substitute needs to act as a binding-agent (e.g. 'banana'), for dessert sometimes whipable egg substitutes (e.g. 'aquafaba') are needed. Another example is cream, where only 'soy cream' is whipable, but the other substitute will leave the user with disappointing results. Still, it is challenging to include such characteristics as features.

C. Future Applications

From a more UX-centric perspective still, our model is a console-based application that is not included in some further application. It remains an open question of how to best include recommendations in yet existing recipe applications and websites. Especially given that ingredients substitution is not only a task of replacing ingredients in a table but includes the adaptation of measurements and instruction texts.

V. Conclusion

In this paper, we have been concerned with the exploration of word embeddings to be used in the domain of recipes and ingredients, especially for the task of substituting ingredients with plant-based equivalents. While this, by itself, offers new perspectives on the topic of plant-based diets that have, not been in focus of sustainable ICT so far, the work is still at its beginning. The various challenges that we briefly discussed remain open for future work.

VI. Poster Presentation

We would like to present our work at the ICT4S conference to show how machine learning algorithms could potentially support plant-based diets. To foster discussions about further optimization of the algorithm we want to provide the technical details and mathematical notations on the poster. Furthermore, potential applications and UX-optimisation should be discussed with a clickable prototype.

References

[1] D. Lawo, K. Litz, C. Gromov, H. Schwärzer, and G. Stevens, "Going vegan: The use of digital media in vegan diet transition," in *Proceedings of Mensch und Computer 2019*, 2019, pp. 503–507.

[2] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Computing Surveys*, vol. 52, no. 5, p. 1–36, Sep 2019.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, p. 3111–3119.

[4] M. Kazama, M. Sugimoto, C. Hosokawa, K. Matsushima, L. R. Varshney, and Y. Ishikawa, "A neural network system for transformation of regional cuisine style," *Frontiers in ICT*, vol. 5, p. 14, Jul 2018.

[5] C. Draschner, J. Lehmann, and H. Jabeen, "Smart chef: Evolving recipes," *EVO* 2019*, p. 8, 2019.

[6] M. De Clercq, M. Stock, B. De Baets, and W. Waegeman, "Data-driven recipe completion using machine learning methods," *Trends in Food Science & Technology*, vol. 49, p. 1–13, Mar 2016.

[7] P. Achananuparp and I. Weber, "Extracting food substitutes from food diary via distributional similarity," *arXiv:1607.08807 [cs]*, Jul 2016, arXiv: 1607.08807. [Online]. Available: http://arxiv.org/abs/1607.08807

[8] P. F. Cueto, M. Roet, and A. Słowik, "Completing partial recipes using item-based collaborative filtering to recommend ingredients," *arXiv preprint arXiv:1907.12380*, 2019.

[3] radimrehurek.com/gensim/