Ingredient Substitution And Filtering Using Natural Language Processing Techniques

Author: Samuel Bowen, C1931367
Project supervisor: Jose Camacho Collados
Module code: CM3203
Number of credits: 40

Project Description

In 2018, approximately seven-million people in the United Kingdom were following a meat free diet (vegetarian, pescaterian and vegan diets), and that number is expected to increase significantly in the coming years[1]. As the number of people following meatless diets increases, it will become important to adapt traditionally carnivorous recipes into meatless recipes by substituting ingredients.

Making international recipes can be difficult for several reasons. International ingredients can be expensive due to travel costs, and can be difficult to source. Ingredient substitution can make international recipes more accessible while minimally altering their flavour profile.

Several papers have shown that natural language processing techniques (primarily word embeddings) can be used to suggest generally applicable ingredient substitutes. One paper uses the Word2Vec and BERT algorithms trained on the 1m+ recipe dataset[2] to produce context-free ingredient substitutions by comparing ingredient similarity [3]. Another paper uses a skip-gram model with negative sampling and compares ingredient similarity [4].

I intend to continue this research by exploring the 1m+ recipe dataset using the BERT algorithm, creating a program that can suggest generally applicable food substitutions and can filter it’s results by dietary restriction(carnivore, pescatarian, vegetarian, vegan).

The filter will be created independently from the ingredient substitution model. I will create a labelled dataset containing all the ingredients within the 1m+ recipe dataset, and each ingredient will be labelled as either carnivorous, pescatarian, vegetarian, or vegan. This dataset will be used to train a classifier. This classifier will be used to filter the results of the ingredient substitution model.

References

https://www.food.gov.uk/sites/default/files/media/document/food-and-you-wave-5-secondary-analysis-current-food-landscape.pdf [1]
https://arxiv.org/abs/1810.06553[2]
https://www.scitepress.org/Papers/2021/102020/102020.pdf [3]
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9094119 [4]

Project Aims and Objectives

I will split my project aims and objectives into two categories: essential and extension. I intend to complete all essential project aims. If I have enough time remaining after I have completed my essential aims and objectives, I will start to complete my extension aims and objectives. Essential aims and objectives are written in black. Extension aims and objectives are written in blue.

- Clean and normalise the 1m+ recipe dataset.
- Use the BERT algorithm to train a model on the 1m+ recipe dataset that will suggest generally applicable ingredient substitutions.
    - Report on the performance of the trained model.
    - Use other word embedding algorithms (Word2Vec, GloVe, fasttext) to train models that can suggest generally applicable ingredient substitutions.
        - Compare the results of all trained models.
- Scrape a labelled dataset from the internet that can be used to create an ingredient classifier (the classifier will identify whether ingredients are vegan, vegetarian, pescatarian, carnivorous)
    - Clean and normalise the scrapped dataset.
- Train a classifier using the scrapped dataset that can be used to filter the results of the ingredient substitution model(s) by dietary requirement. The classifier must be able to accurately classify all ingredients within the 1m+ recipes dataset, but should not be expected to identify ingredients outside of the 1m+ recipe dataset.
    - Report on the performance of the trained classifier.
- Create a GUI for the trained model(s) using Flutter.

Timescale

A gantt chart titled “GANTT Chart C1931367.xlsl” has been attached as a supporting document.

I intend to write my final report whenever I have the time in between implementing my project.

Deliverables and approximate time they should be finished:

- A cleaned and normalised 1m+ recipe dataset. Delivered by the 20/02.

- An ingredient substitution model. Delivered by 06/03.
- A cleaned, normalised and labelled recipe ingredients dataset. Delivered by 20/03.
- An ingredient classifier model. Delivered by 03/04.
- The source code of both. Delivered by 03/04.
- A final report. Delivered by 13/05.

Extended deliverables and approximate time they might be finished:

- Additional ingredient substitution models. Delivered by 24/04.
- The source code of the additional ingredient substitution models. Delivered by 24/04.
- A GUI for the trained models. Delivered by 01/05.

I have scheduled fortnightly meetings with my project supervisor. The approximate dates of my review meetings can be found on the attached gantt chart.