Nesta & BIT Data Science Task

Recipe Ingredient Substitution

Deadline: 4 hours from the time that you receive this task.

Please submit your completed assignment to hr@bi.team. Please also feel free to ask us any questions or clarifications while completing the assignment.

Part 1: Code sample

Please submit the following:

- A sample of your code from an existing project. This could be a short script, repository or notebook. It should be your original work. If it is a collaborative piece then please let us know which part was your contribution.

Please do not spend any time cleaning, tidying or updating your code. The main focus of this assignment and of our assessment will be Part 2. We will not try to run the code. If you do not have a code snippet you are able to share, please get in touch.

Part 2: Data science assignment

Context

Your team is prototyping an app to encourage healthier eating. The aim is to suggest alternative ingredients, with a lower calorie density[1], for online recipes. When a user views a recipe, the tool will highlight ingredients that are contributing the most to the calorie content of the meal and suggest alternatives that will provide a similar culinary result and lower the overall calorie content. You have been tasked with developing a data-driven prototype to power the alternative ingredient suggestion engine of the tool.

The exercise

We would like you to propose a methodology for developing a model or algorithm that uses the data described in the annex below. The result should be able to take features of a recipe and one ingredient it contains as inputs, and then return one or several appropriate alternative ingredients. *Note that it is correct we have not*

[1] A simple definition of calorie density is available here: https://www.nesta.org.uk/report/the-future-of-food-opportunities-to-improve-health-through-reformulation/

provided the actual data, we are interested in your suggested methodology, not the execution of it.

As an example, your model might suggest that the *Cheddar cheese* in a recipe for *cheese on toast* be replaced by *Emmental cheese*.

Please submit the following:

- A short slide-deck summarising the approach that you would take to complete the task:
    - Description of your approach and the methods you would adopt and why
    - Where possible, note specific software libraries and techniques you might use if you were to write the code to tackle this problem
    - Summary of challenges and limitations you might face and how you could tackle them
    - Ideas for future work and development if this were a real-world project
    - The summary should be accessible to a non-expert audience. If you are invited to interview we will ask you to present this for 10 minutes.

Please note:

- We are interested primarily in your proposal for the part of the project that uses the data to suggest sensible alternative ingredients, rather than the calorie reduction aspect. If this were a real-world project, this part could be added easily afterwards. Therefore the data we describe contain no fields for the calorie density of individual ingredients.
- There are different ways to approach this task. There is no 'right answer' or single correct approach. You might choose from a variety of techniques, for example supervised and unsupervised machine learning, natural language processing or network science.
- We also acknowledge the limitations of the data and the timeframe to complete the task. We also understand that real-world data are imperfect and that you will need to make some assumptions to complete this task.

What are we looking for?

- How you make sense of and approach a data science problem and how you would evaluate your results
- Your awareness of the methods that can be used to tackle a data science problem or your ability to research them (we do not expect you to know every possible tool and library that exists)
- Your ability to evaluate the strengths and limitations of different data science approaches

- Your understanding of and ability to anticipate practical challenges that appear in applied data science work, and how to tackle them
- Your ability to communicate your work effectively

Annex: datasets

ingredients.csv

Description: A dataset of ingredients and the food groups that they belong to.
Size: ~10,000 rows
Dictionary:

Field	Type	Description
id	int	A unique ID for each ingredient
name	str	The name of the ingredient
description	str	A one sentence description of the ingredient
food_group	str	One of ~100 high level food categories that the ingredient belongs to. For example, "fruit", "bread", "cured meat", or "oil".

recipes.json

Description: A dataset of recipes from an aggregator website that has agreed to partner with Nesta and BIT to provide data and trial the model. The recipes cover a range of cuisines and diets. There may be different recipes for the same meal.
Size: ~400,000 rows
Dictionary:

Field	Type	Description
id	int	A unique ID for each recipe
name	str	The name of the recipe
description	str	A description of the recipe given by the original author
ingredients	list of int	A list of unique IDs matching ingredients found in ingredients.csv.
steps	list of str	The text describing each step of the recipe.
total_calories	int	Total number of calories the recipe provides.
n_servings	int	Number of servings the recipe provides.

prep_time	int	Time it takes to prepare the meal in minutes.
cook_time	int	Time it takes to cook the meal in minutes.
rating	float	Average rating given by users of the recipe using a 5 star rating system.

replacement_judgements.csv

Description: A set of crowdsourced judgements on the suitability of one ingredient to replace another for a particular recipe. Your colleague obtained these judgements through a crowd sourcing platform. Contributors were given a recipe name, a short description of the recipe, the name of one of the ingredients from the recipe and the name of a suggested replacement for that ingredient. They were then asked to rank the suitability from 1 (not suitable) to 5 (very suitable). Each recipe and replacement was seen by at least 5 contributors and their scores were averaged to obtain a final result.

The dataset uploaded to the platform was created by sampling 8,000 randomly selected recipes (with replacement) and then randomly selecting an ingredient from the recipe to substitute. A suggested replacement was chosen by selecting a random alternative ingredient with the same food_group as defined in ingredients.csv. The users were not asked to take the calorie density of each ingredient into account.

Size: 8,000 rows

Dictionary:

Field	Type	Description
id	int	A unique ID for each recipe and ingredient pair.
recipe_id	int	The unique ID for the recipe shown (matches to recipes.json).
original_ingredient_id	int	The unique ID for the original ingredient in the recipe (matches to ingredients.csv).
replacement_ingredient_id	int	The unique ID for the proposed replacement ingredient (matches to ingredients.csv).
mean_rating	float	The average contributor rating for the suitability of the ingredient replacement. Normalised to a decimal value ranging from 0 to 1.
n_ratings	int	The number of contributors who rated the suitability.