Match Branded Food Items to Common Food IDs in Large Dataset

We have two datasets, both provided as CSV files:

Branded Foods Dataset: Contains 725,000 items, which are all branded foods sold in the United States (e.g., "Sun-Maid California Pitted Prunes 16 oz Canister").
Common Foods Dataset: Contains 4,000 items, which are general food categories (e.g., "prunes").

The task is to link each branded food item in the Branded Foods Dataset to the corresponding ID in the Common Foods Dataset. You will create a new column in the branded foods CSV file called "common_id" (or similar) and populate it with the appropriate common food ID.

Fields provided for each branded food include brand_owner (e.g. Sun-Maid Growers), brand_name (optional), subbrand_name (optional), UPC (barcode), branded_category (e.g. Fruit / Processed), branded_name (e.g. Sun-Maid California Pitted Prunes 16 oz Canister), plus irrelevant fields such as nutrition and serving sizes.

Fields provided for each common food include category (e.g. dried fruits), name (e.g. prunes), and additional serving, expiration, and nutrition information.

Approach: You can use AI or manual methods, as long as the results are reasonably accurate and well-checked. We strongly suggest using AI tools to generate a semi-accurate common_id field, and manually checking the results.

Deliverable: A completed branded foods CSV file with the new "common_id" column filled in for as many items as apply. We estimate that up to 15% of branded foods may not have a reasonable common food match; these can be excluded.

Required skills include experience with large datasets, proficiency in data matching, cleaning, and validation. In addition, you should have familiarity with tools like Python, Excel, or similar. Knowledge of AI and machine learning techniques are necessary to complete this project in a reasonable amount of time.

You will be asked to answer the following questions when submitting a proposal:

1. Please describe, in detail, how you will use AI to generate a reasonably accurate common_id field, given the provided fields in both the branded foods and common foods datasets.
2. Assuming you are using an AI tool to generate the initial common_id values, what strategies will you use to check the results? How long do you foresee this taking?
3. How will you determine which branded foods do not have a reasonable common food match? These ingredients should have a blank common_id field.