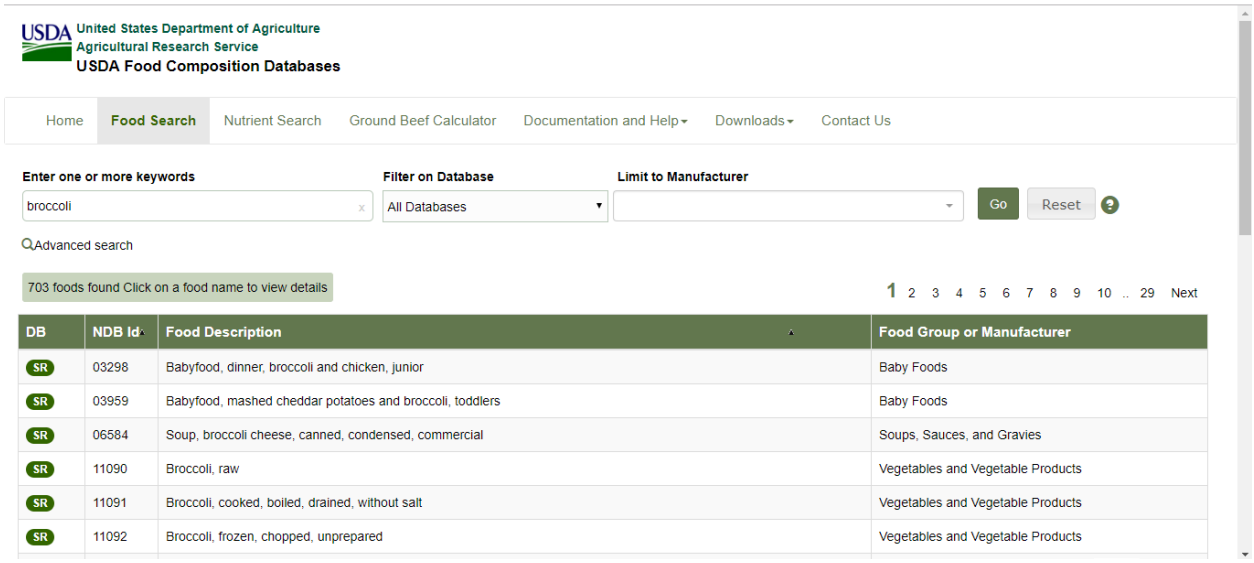
2019/02/07

Querying Ingredients with USDA Database

Introduction

Given a food ingredient, we want to know the amount of nutrients each serving of the ingredients have. To do so, we query the USDA database.
https://ndb.nal.usda.gov/ndb/search/list?home=true
However, there is a problem, as shown in the figure below. Each query returns multiple results and not all of them are relevant. Our goal is to find the most relevant result (NDB Id). In our example, 11090



DB	NDB Id	Food Description	Food Group or Manufacturer
SR	03298	Babyfood, dinner, broccoli and chicken, junior	Baby Foods
SR	03959	Babyfood, mashed cheddar potatoes and broccoli, toddlers	Baby Foods
SR	06584	Soup, broccoli cheese, canned, condensed, commercial	Soups, Sauces, and Gravies
SR	11090	Broccoli, raw	Vegetables and Vegetable Products
SR	11091	Broccoli, cooked, boiled, drained, without salt	Vegetables and Vegetable Products
SR	11092	Broccoli, frozen, chopped, unprepared	Vegetables and Vegetable Products

Problem Formulation

Input: a set of keywords describing a food ingredient, about 2.5k queries in total
Output: The most relevant NDB Id among the results the database returns

Method

To find the most relevant result, we **find the smallest Id in the correct food group (this needs to be checked).** The food groups are: Dairy and Egg Products; Spices and Herbs; Baby Foods; Fats and Oils; Poultry Products; Soups, Sauces and Gravies; Sausages and Luncheon Meats; Breakfast Cereals; Fruits and Fruit Juices; Pork Products; Vegetables and Vegetable Products; Nut and Seed Products; Beef Products; Beverages; Finfish and Shellfish Products, Legumes

and Legume Products; Lamb, Veal and Game Products; Baked Products; Sweets; Cereal Grains and Pasta; Fast Foods; Meals, Entrees and Side Dishes; Snacks; American Indian/Alaska Native Foods; Restaurant Foods.

The problem then becomes determining the correct food group of each query. To do so, we compute the distribution (frequencies) of the food groups returned by the database.

- If the most frequent food group dominates, then we choose it as the food group
- If there are more than one dominated food groups, we have to choose one of the dominated groups manually

Dealing with failure cases:

1. If the initial query (a set of keywords describing a food ingredient) returns nothing, one suggestion is to query with only those nouns (check NLTK for help).
2. If above query still returns nothing, we record this ingredient and query it manually.