Ingredients Prediction Project

Neil Patel (neipatel@illinois.edu)

December 2019

1 Introduction

We have faced the daunting task of eating healthier, but the convoluted and unorganized information makes this task inconvenient and unbearable. Dietitian/researcher focus on different information to guide their customers. The information could involve calories, carbs, saturated fats, etc. This information is standardized and it is displayed on all the items sold in the USA. However, the nutrition facts do not tell the full story of the product.

The ingredients themselves are more important than the nutrition facts. Many of us know that Avocado is a super food, but the nutrition facts might confuse our judgement, see figure 1. The 29 grams of fat might be misleading, as most of it is unsaturated fats and it is easy for fad dieters to over look the fiber and vitamins it provides. Even some of the nutrition scoring systems around the world will under estimate an avocado because these systems only use the nutrition facts. Despite the difficulties of evaluating an avocado, evaluating a **"single well known"** item is much easier than evaluation bread, ice cream, snacks, etc. This project will score the products by their ingredients using information retrieval and machine learning techniques.

Nutrition Facts

Avocados

Amount Per 1 avocado, NS as to Florida or California (201 g)			
Calories 322			
			% Daily Value*
Total Fat 29 g			44%
Saturated fat 4.3 g			21%
Polyunsaturated fat 3.7 g			
Monounsaturated fat 20 g			
Cholesterol 0 mg			0%
Sodium 14 mg			0%
Potassium 975 mg			27%
Total Carbohydrate 17 g			5%
Dietary fiber 13 g			52%
Sugar 1.3 g			
Protein 4 g			8%
Vitamin A	5%	Vitamin C	33%
Calcium	2%	Iron	6%
Vitamin D	0%	Vitamin B-6	25%
Cobalamin	0%	Magnesium	14%

Figure 1: Avocado Nutrition Facts

2 Novelty

2.1 Related Work

Ingredients aren't a target for most dietary assistance application. There are many popular applications that focus on calorie counts and the nutritional table: MyFitnessPal, Lose It!, MyNetDiary. These applications are focused on weight-loss, and tracking calorie consumption.

Open Food Facts [1], is an open source MongoDB Database with over one million scanned items. It provides a list of ingredients by scanning the barcode of popular items. The database is also open for new additions. Open Food Facts does some analysis on the ingredients and nutritional facts, but they are rudimentary. This is the part of their analysis that could be improved by IR and AI.

Unlike other projects in this discipline, this project will use ingredients to evaluate the healthiness of a product. The main questions that this project is trying to answer: "Is this product

healthy?".

2.2 Methodology

The high-level idea is to use the correlation between ingredients and nutrition facts to better estimate the quality of the products. Products that have ingredients closer to healthy items (e.g. whole wheat) will get a better score versus products that have ingredients that relate to unhealthy products (e.g. sugar, sweeteners)

Since this system will only require a query, it does not need the manual effort of inputting product data. Products are changing constantly and this system will be able to score products regardless of the items presents in the database.

2.3 Implementation

2.3.1 Ingredients/Products

The project started with a search for a database that had enough information to correlate ingredients to nutrition facts. There is an open source database that is available through Open Food Facts [1]. The database is in MongoDB, and the dump of the database is uploaded periodically. I was able to obtain the dump from the their website. This database was very valuable, as it had over 1 million products. Like many data sources, their was some cleaning that was required to obtain a more successful correlation.

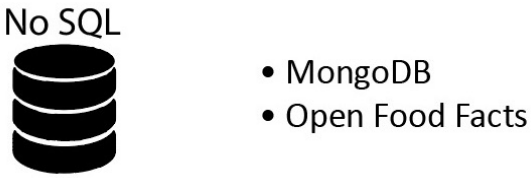

Figure 2: Ingredient/Product Database

2.3.2 Data Cleaning

The products from the database were extracted to a local database. There were some strict requirements. The product document must include: fats, saturated fats, carbohydrates, fiber, sugar, protein, calories. The ingredients should also be provided in English. After going through this strict cleaning process, there were only 117,398 items that were obtained from 1,081,561. This was less than 11 percent of the database. Each item also was scored based on the nutrition facts only.

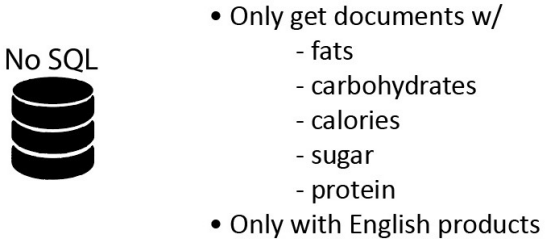

Figure 3: Data Cleaning

2.3.3 Training/Extra Data

Since most products have multiple ingredients, I added some training/extra data that will help evaluate the system and assist in the correlation results. Products such as vegetables and fruits are not sold as packaged items and rarely have nutrition facts. Nuts, and other high fat items tend to get lower scoring in a nutrition fact score due to the high amount of (healthy) fats. To help the system, I experimented with this extra data.

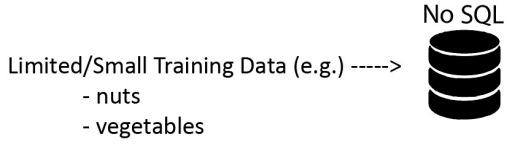

Figure 4: Training

2.3.4 Stop-Word

Stop words in common literature is well-known but those same words are not that common in ingredients data. After having a quality dataset, I obtained the most frequent words in the database. Unfortunately, the most frequent words were not good stop-words, and they had a lot of meaning. The most frequent words are listed. Those ingredients are some of the worst ingredients in the food

that we eat and provide a lot of meaning to the scoring.

Top 10 ingredients
salt
sugar
flavouring
dairy
vegetable oil and fat
cereal
oil
vegetable oil
wheat

Instead of the most common words, I looked at ingredients that were pairs or triplets, like "sea salt" or "California wild rice". Many times, the first word is an adjective and I was able to include it in the stop-word list.

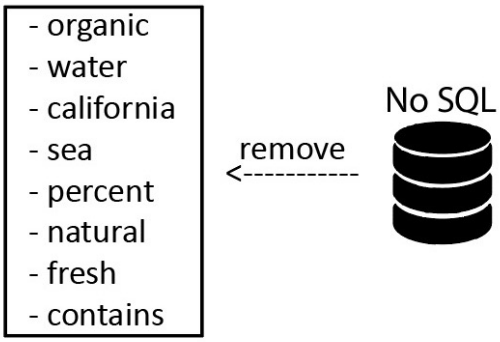

Figure 5: Stop-Words

2.3.5 Lucene Index/Search

For the document/ingredient search, I used Lucene and the BM25Similarity for searching. The query went through the same string cleaning as the database steps earlier. I tried to experiment with the hierarchy of the ingredients but I was not able to quantify the results. I believe giving the structure to the ingredients by quantity will greatly affect the scoring results.

2.3.6 Scoring Algorithm

After the search results, the top 25 hits were used to evaluate the healthiness of the matches. Each

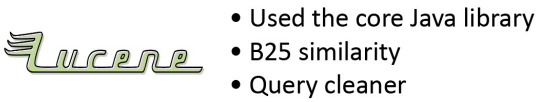

Figure 6: Lucene Index/Search

product in the database was indexed with a score (from the trivial French Nutrition scoring system). A minimum of 2 results would be required to score the ingredients query. The hit score was taken on a logarithmic scale to let lesser matched documents to provide scoring. The results were shifted to a 0 to 100 scale, where the higher the value the healthier the item.

- Logarithmic
- Top 25 matches use
- Minimum of 2 matches to rate
- 0-100 point system

*(B25_score/norm) * ((nutrition score - 12.5) * -27.5) * 100/55*

Figure 7: Scoring Algorithm

2.3.7 Web Application

A web application was created for supermarket customer to input the ingredient data and get results while shopping. The application is able to output the result almost instantly.

2.4 Informative Failures

The most notable failure was using deep learning to find a correlation between the ingredients and the nutrition facts. I used Tensor Flow to create a network, where the inputs were the ingredients and the output was the score. The input vocabulary was converted to decimal values, which were spread evenly between 0 and 1. After using 80 percent of the database, I was not able to get a better correlation than the Lucene document search results.

The results are great for comparing related items, but it is not valuable for unrelated items and it is also not valuable for short list of ingredients. This

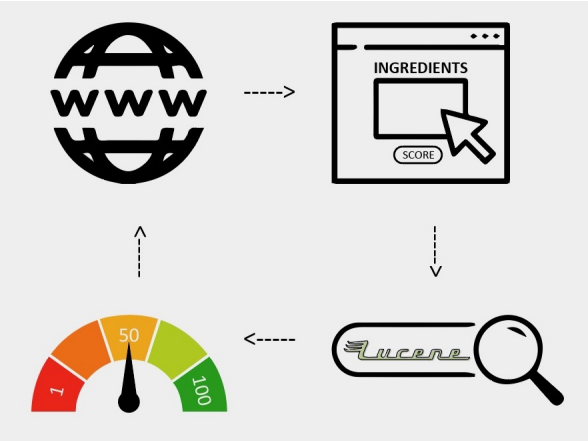

Figure 8: Web App

is the part of the algorithm that will need to imporve the most.

2.5 Future

Other than improving the algorithm, there are some other areas that can help the user interaction. Moving this project from a web application to the mobile application will be more suited for the supermarket. Using an OCR library to read ingredients will help the user search long list of ingredients quicker. A nutrition would also be able to provide more background on many of the food items. Having expert feedback could provide better scoring expectations.

3 Test

Testing and creating a bench mark was very difficult for this problem. There is no system to rate ingredients. Even nutritionist and doctors will disagree on the healthiness of a particular ingredient. To test the algorithm, I had to use clearly known unhealthy/healthy ingredients to compare the results. Some of the results from testing are shared in this report but more are included in the demo.

3.1 Bread and Ice Cream Results

The breads in figure 9 are best rated breads by the French nutrition algorithm. In two of the breads, there many added sugars that are found in junk food. The two ice creams that are being compared are drastically different in healthiness and the algorithm is able to indicate the difference.

Ingredients	French Nutrition Scorer	New Scorer
ingredient cracked wheat powerseed mix wheat gluten gluten fruit juice fruit oat fiber sea salt cultured wheat yeast vinegar wheat cereal wheat	87.3	77.0
wheat flour cereal wheat flour cereal flour wheat gluten gluten high fructose corn syrup glucose fructose corn syrup glucose fructose syrup soya bean soya	87.3	74.1
wheat flour filtered sugar salt yeast e282	83.6	77.0
unbromated unbleached enriched wheat flour oat cereal high fructose corn syrup glucose fructose corn syrup glucose fructose syrup yeast soya oil oil vegetable oil fat vegetable oil	83.6	67.9

Item	Ingredients	Score
Healthy Ice Cream	avocado puree avocado oil cane sugar tapioca starch cocoa powder vanilla extract sea salt guar gum gun acacia	61.6
Unhealthy Ice Cream	milk sugar cream chocolate dextrose milk fat cocoa butter carob gum vanilla extract natural flavor chocolate processed alkali soy lecithin mint leaf extractives	34.2

Figure 9: Results

4 Demo/Documentation

Website and demo instructions are at: https://github.com/nbp828/IngredientsSearchJava

The web application is very simple to use. There are two pages. The home page is a form to input your ingredients. The ingredients need to be separated by commas. Once the user pressed the score button, the algorithm will search for relevant documents that match the input data.

The range of the score is from 0 to 100. The higher the score, the better the food item. The score is a more relative than a great comparison between very unrelated products. For example, kale 74.0 is black beans is 85.3. However, milk is 70.1, and cream is 51.0.

For most products, there isn't just one ingredient. Despite this, this will be the next area to improve. Some experiments that I started was to normalize the score by category and use the ingredients hierarchy (ingredients are displayed in-order of quantity).

The application currently is already helpful to take to the supermarket. It is able to rate complicated ingredients list. Due to the querying/search nature of the implementation, it is reasonable to get some poor results for rare data. In the example with kale,it is likely that the search is mostly matching granola bars that have little kale in it.

Since this is the only search engine and scorer for ingredients, there are many directions to improve this application.

5 Repository

Description	Repository
Python/Tensorflow	https://github.com/nbp828/FoodPyExample
Java/Lucene	https://github.com/nbp828/IngredientsSearchJava
Web App	https://github.com/nbp828/webapp2

References

[1] Open food facts.