A Programmatic Approach to Parsing Ingredient Lists from Consumer Packaged Goods for Effective Data Analysis

Molly Hirsh, Edward Appiah, Shirley Mach, Lauren Zhovmer, Kasey Heintz
U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition

Abstract

INTRODUCTION: Advancing technology to increase usability of ingredient list information from consumer packaged goods is of interest to both the government and private sector. Enhancing parsing technology allows full-text ingredient labels to be divided into individual ingredient terms, while eliminating extraneous noise, preserving inter-ingredient relationships, and capturing ingredient predominance in each product.
METHODS: Food label data was queried from products in FDA's structured food database, FoodTrak. Products were limited to seven food categories and excluded if they had no barcode identifier or ingredient list. Iterative steps were designed in structured query language (SQL) to parse full-text ingredient labels while maintaining their relational hierarchy. These queries transformed ingredient lists containing various separators into exclusively comma separated lists for parsing in a layered approach. Three tables of specific terms– the REMOVE registry, CONVERT registry, and PRESERVE registry- were manually created during logic development to update queries and mitigate terms that did not adhere to the comma-based parsing structure. RESULTS: 204,982 product records qualified for parsing. Products were categorized as bread (16%), condiments (10%), cookies (18%), crackers (8%), frozen meals (18%), ice cream (19%), and soup (11%). Parsing resulted in 96,854 unique ingredient terms, with an average 18,854 ingredient terms per category.
DISCUSSION/CONCLUSION: The resulting parsing technology enabled users to quickly query products containing a specific ingredient, identify related ingredient terms, and view contextual descriptions of the ingredient on the label. Applying this parsing technology has many applications in big data. Future goals include utilizing the resulting unique list of parsed terms to build a comprehensive food label ingredient thesaurus from synonymous terms, find connections between co-occurring ingredient terms, and contribute towards effective post-market ingredient analyses while supporting other FDA databases.

Introduction

The ingredients in the US food supply are evolving with innovation in food technology and changing food landscapes. Assembling these ingredients in an organized database is useful in many capacities for the FDA. For example, an ingredient database can be used to quickly pull all relevant products containing an ingredient of interest, to evaluate ingredient prevalence, or to help inform key elements of FDA's Nutrition Innovation Strategy.[1] Availability of internal resources to parse ingredient labels offers advantages including standardization of data and developing comprehensive approaches for food monitoring applications, while averting external ingredient label parsing inconsistencies and gaps in coverage.
To parse ingredient terms, considerations should be made to remove noise in ingredient labels: non-ingredient phrases such as product claims (e.g., "With 10% Less Fat than Regular Ice Cream"), conditional statements (e.g., "One or More of the Following"), or quantifying phrases (e.g., "Contains Less than 2% of"). Care should be taken to avoid separating related terms split by conjunctions (e.g., Mono and diglycerides). Additionally, effort to maintain meta-data associated with ingredient lists, such as the hierarchical relationship between terms and sub-ingredients, ingredient order of predominance, and general co-occurrence, can expand future predictive analytic capabilities. The technology described in this poster can successfully parse full text ingredient labels from multiple database sources into individual terms for optimal ingredient searching efficiency.

"This project was supported in part by an appointment to the Research Participation Program at the U.S. Food and Drug Administration administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration."

Materials and Methods

Ingredient label data was obtained from FDA's structured food database (FoodTrak), which includes all available data from Mintel, Label Insight, Syndigo, and NuVal for consumer packaged goods collected from 1996 to 2020. Product records were standardized between the multiple data sources and imported into Microsoft SQL Server 2016.
A unique ID was assigned to each product record. A table of full-text ingredient lists, record IDs, Universal Product Codes (UPCs), and category names was extracted from the full dataset. All historical product records related to the following seven categories were utilized for ingredient parsing: Breads, Cookies, Crackers, Frozen Meals, Ice Creams, Soups, and Condiments. Products without ingredient labels were excluded.
After examining patterns within the raw ingredient lists, a list of delimiters (words or symbols that come between two ingredient terms) was logged. Next, ingredient terms that did not conform to delimiter logic were identified and added to registries (Tables 1-4). Each registry served a different functionality to reformat the terms: the CONVERT registry contained terms that required manipulation to avoid improper separation during parsing; the REMOVE registry contained non-food related noise meant to be cleared from the parsed ingredient output; the PRESERVE registry contained terms that inherently included known delimiters, which were exempted during the parsing process. Registry terms were initially identified manually. As additional products were parsed, automated logic was applied to expand information contained in the registries.
As shown in Figure 1, ingredient text was prepared for parsing, first by grouping related sub-ingredient lists. Terms listed either after a colon or between brackets were flanked with parenthesis. Where text contained only one opening or closing parenthesis, the punctuation was replaced with commas. Ingredient text was then scanned and manipulated based on any matches encountered in the CONVERT and REMOVE registries.
An iterative process was used to parse ingredients and related sub-ingredients (those grouped in parentheses) while indexing the positional relationship of each term in the overall list. The following changes were applied only to terms outside parenthesis. The beginning and end of each term was trimmed of spaces and delimiters. Ingredients were then parsed by delimiters. When parsing, terms encountered from the PRESERVE registry were skipped. If a parenthesis flanked group remained, it was separated and stripped of the leading and trailing parenthesis. The iterative portion of the process repeated layer by layer until no delimiters were detected in the final parsed ingredient list. As a final cleanup, all blank cells and extra spaces were removed; ingredient terms were ordered as they originally appeared on the ingredient panel; and terms with subsequent sub-ingredient lists were marked with an indicator called 'COMPOUND'.

Convert Registry

CONVERT FROM	CONVERT TO
GUMS (XANTHAN, CELLULOSE)	GUMS (XANTHAN GUM, CELLULOSE GUM)
MONO- AND DIGLYCERIDES	MONOGLYCERIDES, DIGLYCERIDES
PALM, SOY, AND/OR CANOLA OIL	PALM OIL, SOY OIL, CANOLA OIL
TOMATO CALCIUM SULFATE	TOMATO, CALCIUM SULFATE

Remove Registry

REMOVE
CONTAINS LESS THAN 2% OF EACH OF THE FOLLOWING
MADE WITH SMILES
AND LOTS OF LOVE

Preserve Registry

PRESERVE	EXEMPT CLAUSE
HALF&HALF	&
FD&C	&
L. BACILLI	.
[0-9].[0-9]	.

Delimiters

DELIMITERS		
AND	AND/ OR	-
OR	AND /OR	;
AND/OR	WITH	.
AND / OR	:	&

Tables 1-4. Ingredient Parsing Registry and Delimiter Examples

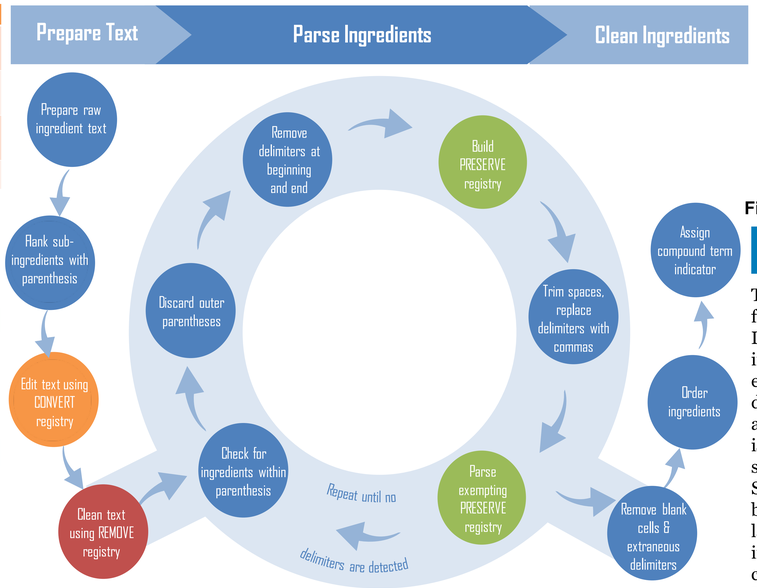

Figure 1. Ingredient Parsing Process Overview

Results

Of the seven categories chosen for analysis, 110,538 distinct products with 204,982 total records qualified for ingredient parsing. Product records were distributed as follows: bread (n=31,861), condiment (n=21,170), cookie (n=37,655), cracker (n=17,105), frozen food (n=35,947), ice cream (n=37,982), and soup (n=23,262).
The CONVERT, REMOVE, and PRESERVE registries housed 3,185, 1,382, and 26,800 terms, respectively.
A parsed ingredient label sample is shown in Figure 2. A total of 6,009,049 ingredient terms were parsed from these product records, 96,847 of which were unique. The category average was 18,854 unique ingredient terms, with a breakdown of ingredient terms (Figure 3) as follows: n=15,018 in bread, n=12,721 in condiments, n=22,078 in cookies, n=10,582 in crackers, n=38,505 in frozen food, n=17,972 in ice cream and n=15,108 in soup.

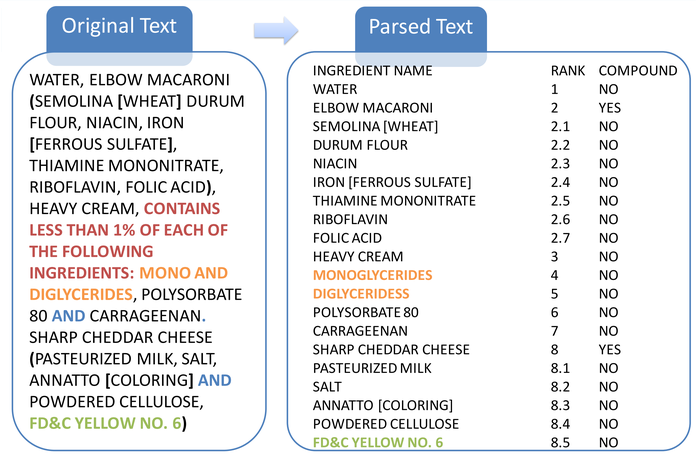

Figure 2. Ingredient Raw Text to Parsed Text Example

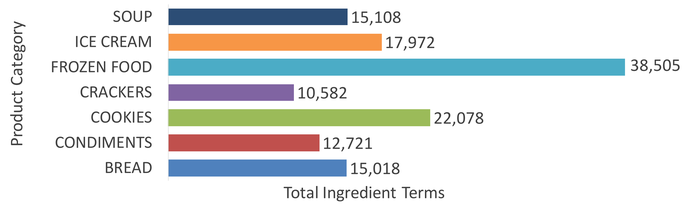

Figure 3. Count of Ingredient Terms in Each Product Category

Discussion and Conclusion

This ingredient parsing technology facilitates the extraction of ingredient information from product-label data while preserving the relative predominance of each ingredient. It also preserves the relationship between compound ingredients and their sub-ingredients. Cleaning and parsing the ingredient lists enables more accurate and efficient data functionality such as transforming data into ingredient prevalence dashboards, conducting analyses for monitoring purposes, or constructing queries to address ingredient specific research questions. For example, the term "TAHINI" which is a ground sesame paste, could be searched to help identify products that contain sesame but may not currently list sesame as an allergen.[2]
Some limitations of the technology include the necessity for manual resources in building the ingredient cleaning registries and in making corrections to manufacturer labels that contain mistakes. Additionally, some of the unique parsed ingredient terms included spelling variations and synonyms. An associated synonym database is being created to group these terms for more efficient querying.
The seven product categories were chosen to provide expansive coverage of possible ingredient terms in the total food supply. Future directions include applying the parsing method to other food categories to ensure the developed logic can process ingredient lists not yet encountered and expanding ingredient cleaning registries with natural language processing tools to reduce manual input. The database will be expanded over time as new products are introduced to the market.

1. FDA. FDA Nutrition Innovation Strategy. 2021; https://www.fda.gov/food/food-labeling-nutrition/fda-nutrition-innovation-strategy. Accessed April 19, 2021.
2. FDA. FDA Encourages Manufacturers to Clearly Declare All Uses of Sesame in Ingredient List on Food Labels. 2020; https://www.fda.gov/news-events/press-announcements/fda-encourages-manufacturers-clearly-declare-all-uses-sesame-ingredient-list-food-labels. Accessed April 20, 2021.