Finding Nutritious Alternatives to Ingredients in Recipes Using Machine Learning

Saachi Subramaniam
South Brunswick High School; South Brunswick, United States
Email: saachisubramaniam@gmail.com

Abstract – In the United States, close to 678,000 deaths per year are attributed to nutrition-related causes with obesity and malnutrition as the leading cause of death. By helping users reform their diets, malnutrition in all forms can be combated. This research discusses the factors involved in food nutrition and the integration of nutritious foods into users' lifestyles. Through the utilization of Python, natural language processing and APIs, this work supplements individuals with nutritional alternatives. In this work, Each ingredient is categorized, and analyzed for nutritional value, then a similar product with higher nutritional value is reported to the user. The work also implements a graphical user interface that allows users to easily interact with the database and obtain results quickly.

Key Words – Healthy Eating, Nutrition, Diets, Food Pairings

INTRODUCTION

According to a study done by Pew Research, 72% of American's believe good nutrition is the key to a long and healthy life. However when the same study asked subjects to identify the statement which best describes their eating habits. Only 18% reported their main focus as "health and nutrition." [1] Furthermore, more than 48% of Americans rarely or never seek out information about how their food is produced according to a food literacy study done by Michigan State University. [2] The American perception of food and nutrition is based largely on misinformation and "fad diets." Fad diets refer to specific eating restrictions that become popular for a short time. Often, fad diets are not backed by scientific research but instead, merely promote disordered eating. Many fad diets revolve around cutting out certain food groups and restricting caloric intake. This work proposes an intuitive approach to nutrition rather than a restrictive approach. The work focuses on the nutritional composition of each ingredient whereas fad diet models may focus on certain food groups such as vegetables while excluding food groups like carbohydrates. However, this research was conducted with the intent of incorporating more nutritious foods into users' diets while preserving the taste of dishes, focusing on an intuitive model as opposed to a restrictive model. Some of the benefits of intuitive eating include developing a healthier relationship with food, becoming more attuned to the body's natural hunger and fullness cues, and leading a happier lifestyle. This research's foundations in intuitive eating promotes healthy relationships with food by encouraging users to have awareness for their eating choices. Furthermore, the work gives users the freedom to make their own dietary choices while simultaneously offering nutritious alternatives.

BACKGROUND

This research employed three different tools in order to classify, analyze, and suggest ingredient alternatives. These three tools are Web Scraping, Natural Language Processing and APIs.

I. Web Scraping

This research employed Web Scraping in order to retrieve results from websites such as Google Shopping. Web scraping is the process by which computers extract data from websites. With web scraping, computers send requests to websites, and extract information relevant to the user. Information can then be used for a variety of business and personal related outcomes. The automation of data extraction allows for companies to retrieve competitor's prices, users to receive up-to-date news alerts, and for fantasy sports enthusiasts to calculate the statistics of their favorite fantasy sports leagues. While web scraping is primarily used in price monitoring, market research, and content monitoring, this work utilized web scraping to extract product information from Google Shopping, as well as gather ingredient information from the web.

II. Natural Language Processing (NLP)

Natural Language Processing is a subsection of a larger field of study known as Machine Learning. Machine Learning algorithms identify patterns in training data which allows the algorithms to apply the same patterns to new data. Training data refers to datasets that machine learning algorithms use to learn and identify patterns. When given a new data point, these algorithms are able to use previously identified patterns to draw new conclusions about the data. Machine Learning powers social media feeds, movie recommendation services, and many smart assistants such as Siri or Alexa. Natural Language Processing (NLP) uses the same pattern recognition principle and applies it to oral and written language. NLP is able to

manipulate, analyze, and derive meaning from human language. Natural Language Processing technology has been used in spell-checking devices and spam inbox filtering services. This work implements NLP to categorize ingredients into different food groups based on name.

III. APIs (Spoonacular)

Application Programming Interface (API) is the intermediary connecting two applications together. APIs can perform certain tasks and retrieve information through a user request, and send the information back to the user, or complete the task. Currently, APIs exist for many different fields and topics such as food, music, news and weather. For this work, the specific API used was Spoonacular. The Spoonacular API holds nutrition information for ingredients, recipes, and products. This information was later used in the classification model and in the algorithms for determining nutritional value.

RELATED WORKS

This section discusses similar, pre-existing works related to nutrition, dietetics, and the use of natural language processing in food analysis.

Due to the fact that the concept of a proper diet is rapidly evolving, many researchers have created algorithms for making healthy foods more accessible. Benjamin Share, James Ordner, and Zack Cinquini's "Optimizing meals under dietary constraints" presents the implementation of machine learning algorithms in order to both predict the success of a recipe according to users' dietary restrictions, and make substitution recommendations accordingly. [3] The latter half of the paper discussing methods for food recommendation systems proved most relevant for the purposes of this research. The system analyzed the ingredient to be substituted and the proposed substitution in relation to how often the two ingredients appeared in recipes together. The paper notes that this algorithm worked well for rarer ingredients, however failed to produce accurate recommendations for common ingredients which appear in many recipes. For common ingredients, the work utilized Naive Bayes to construct a more complex approach.

The second related work by Stanley C. Rak, describes a method for determining the nutritional value of a food. [4] Rak used a weighting system to score certain vitamins and nutrients depending on their healthiness. Then, Rak calculated the nutritional value of the food according to the amount of different vitamins, nutrients, and fats present in the food. This approach inspired the conditional algorithm for determining nutritional value.

RESEARCH QUESTION

How can we help Americans make healthier eating choices by providing healthy substitutions for unhealthy food items that may be in their recipes?

APPROACH

This approach contains three different elements within the work. First, an ingredient is classified into a general food group through the classification model. Then, a search term is generated for the ingredient based on the nutrition recommendations for ingredient's classification. Finally, the search term is inputted into Google Shopping where the work scrapes the results and displays them using a graphical interface. Figure 1 above documents the flow of this approach, and briefly outlines each element of the method. *(See Figure 1)*

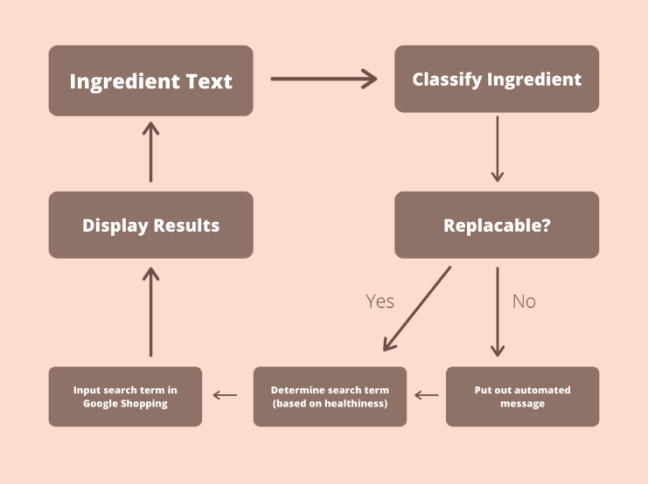

FIGURE 1: Workflow chart for method

I. Classification

This work applied Natural Language Processing and APIs in order to increase the accuracy of the categorization model. The categorization model decomposed ingredients into a general food group by feeding the inputted ingredient through two stages of categorization. Stage one integrated NLP's "Bag of Words" in order to compute a preliminary set of categorizations. The Bag of Words model is a specific approach to Natural Language Processing that converts text into vectors, which computers are better able to comprehend. This model can count the frequency of a set of characters in a given input, and assign a value to the characters accordingly. This allows the model to draw patterns based on the frequency of a set of characters. The "Bag of Words" approach uses sets of training data to teach machine learning algorithms how to interpret new data. The program uses training data to identify key ingredients in each item, and uses this information to sort the food into its respective food group. These food groups are based on the Eurocode Main Food Group Rules. [5] The first stage of classification identified all milks, juices, and common meat products into their respective food categories. For example, when "apple juice" undergoes the first stage of categorization, it is identified as a beverage. The machine learning algorithm uses patterns from the training data to recognize that all juices should be classified as a beverage. All items that are not categorized in the first stage, move to

the second stage. In the second stage, the work utilizes the Spoonacular API to retrieve the grocery store aisle of the inputted ingredient. Each grocery store aisle corresponded with a general food group, therefore each food item's grocery store aisle got converted into a food group. Finally, if an item cannot be classified into a food group, the work notifies the user, and asks for another ingredient.

II. Determining Nutritional Value

This component of the work used the food group from the classification model in order to create search terms for healthier alternatives. Each food group contained different specifications and algorithms for determining nutritional value. These algorithms are as follows: Replacement, Web Scraping, and Conditional. These algorithms take into account the requirements for healthy ingredients, and devise a search term for the ingredient according to the specifications. The replacement algorithm used a list of nutritious and innutritious ingredients to determine the nutritional value of the item. If the item was in the list of innutritious ingredients, the search term for the ingredient used one of the ingredients in the nutritious list. If the item was nutritious, the work notified the user that the proposed ingredient is already healthy. The web scraping algorithm used data from the web to determine the nutritional value of the proposed ingredient. Using the data, the web scraping algorithm produced a search term for the ingredient. The fish category is a notable example of the web scraping algorithm. The web scraping algorithm for the the fish category scraped data from the United States Environmental Protection Agency's mercury levels in fish chart. [6] This chart identified the mercury levels in common seafood products and made recommendations based on the amount of mercury in a fish. The algorithm scraped the website data for the name of the seafood product and the EPA recommendation. Then, a list of safe seafood products was created. If the proposed ingredient had a positive EPA recommendation, the work would notify the user that the proposed ingredient is healthy. If the ingredient had a negative EPA recommendation, the search term for the proposed ingredient would be derived from the list of safe seafood products. The conditional algorithm is the final method used for determining nutritiousness. This algorithm applies conditions to the ingredient, which forms the search term. For example, the grains category included the conditions "whole wheat" because whole wheat grain products are more nutritious than other grain products. "Whole wheat" was then added to the ingredient name in order to create the full search term. Due to the fact that each food category's specifications were of various levels of complexity, some search terms and conditions were harder to identify than others.

III. Web Scraping Google Shopping

After identifying the different categories of foods and the different qualities that made them healthy, the third element of the work transformed the search terms into ingredients that could be substituted into a dish. The process of converting theoretical keywords into real ingredients occurred through Google Shopping. By entering keywords into Google Shopping and web scraping the results, the work was able to bridge the gap between search terms such as "whole wheat pasta" and the names of real ingredients such as "Great Value Whole Wheat Spaghetti." In this component, the work entered keywords into the Google Shopping website, scraped information about the resulting items, and displayed the information in the output. One of the main benefits to using Google Shopping as opposed to other online shopping sites is its accessibility. Google is one of the most universal corporations, with branches across the globe. In order to make results most practical for users, the work uses Google Shopping because it is universal, and contains millions of different ingredients from many different countries. Google Shopping is not limited by geographical location, therefore ingredients are most likely to be very widely accessible. Not only this, but ingredients from countries outside the Unites States are available because Google Shopping is not limited by geographical location.

a. Displaying Products with Regex

In order to make information more legible to users, the work utilized Regular Expressions. Regular Expressions (RegEx) are sequences of characters used as a search pattern. These "capture groups" can be used to delete, replace, and match certain parts of strings. In this work, the web-scraped information contained HTML tags and irrelevant website information. Regular Expressions matched patterns of unneeded characters and deleted them. This converted hard to decipher information into simple and clear ingredients which could then be displayed to users.

IV. Graphical User Interface

This research implemented a graphical user interface which coherently displayed ingredient information to users. The work employed PyGame in order to add functionality to the display. Each page design was imported into PyGame from Figma. Then, buttons and text fields were placed as needed throughout the display. Creating buttons and text fields in PyGame required coordinate locations. Once coordinate locations were provided, users was able to click with the provided coordinates and perform an action. The full graphical user interface features a log in page, sign up page, and home screen where users can enter ingredients and interact with the software. *(See Figure 2)*

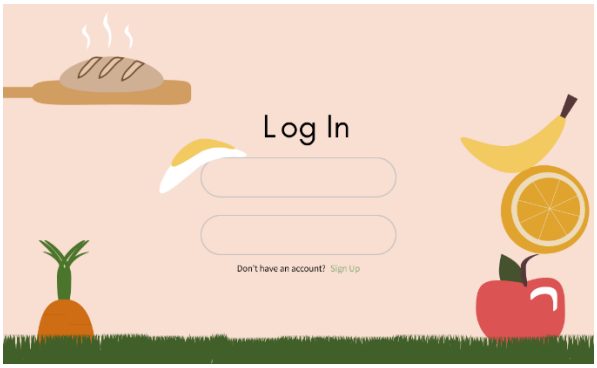

FIGURE 2: Log In page of graphical user interface

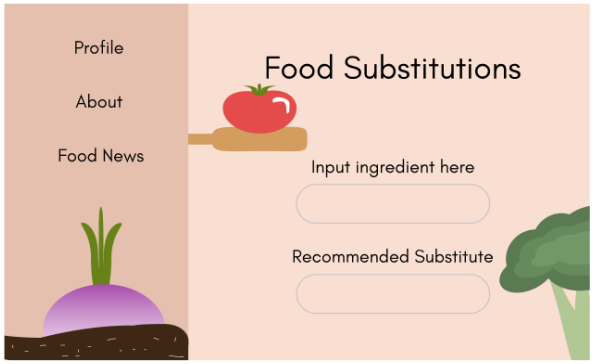

FIGURE 3: Homepage of graphical user interface

RESULTS

This research aimed to integrate healthy choices into users' lifestyles by producing nutritious alternative to potentially unhealthy ingredients. The results of this work produced five nutritious alternatives for the ingredient inputted. This model was able to accurately suggest ingredient alternatives for all fruit, vegetable, protein, dairy, and beverage products, including common household ingredients. However, certain name-brand confectionary items were not able undergo the categorization model successfully. In the future, the usage of multiple datasets may improve the categorization model by allowing it to categorize a wider variety of products. The applicability of these results were then tested by feeding different ingredients with commonly known healthier substitutes. The results were compared against the substitute in order to determine efficacy. Examples *(See Figure 4)* of potential user inputs and results are shown below. Furthermore, the usage of a browser interface allowed each result to be verified in real time. The interface, as opposed to a headless browser, displays the work inputting and retrieving the results through the browser. *(See Figure 5)* The main limitation of this model is the incompatibility with newer versions of Google Shopping. Due to the web scraping component of this work relying heavily on the site design of Google Shopping, as Google Shopping is updated, the web crawler is unable to gather data using the newer version.

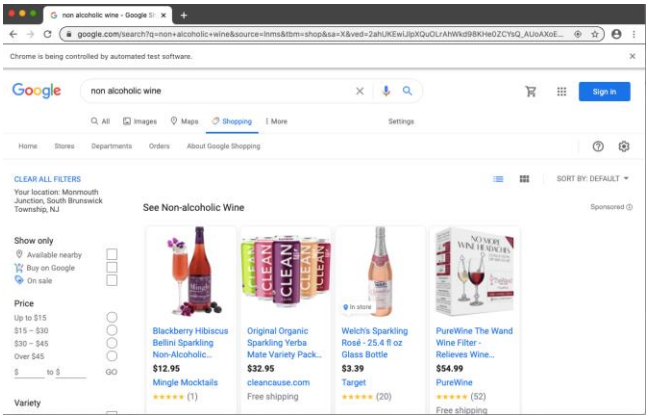

FIGURE 4: Browser interface with sample input non-alcoholic wine

Whole Milk	Silk Soymilk, Original - 64 fl oz Silk Organic Soymilk, Organic, Unsweet - 64 fl oz Silk Original Soy Milk 32 oz Great Value Soy Milk, Original - 0.5 gl (1.89 lt) Silk Soymilk, Very Vanilla - 12 pack, 8 fl oz containers
Venison	Mori-nu Silken Extra Firm Tofu 12.3oz House Foods Tofu Firm, 16.0 oz, Size: 16 fl oz Nasoya Organic Extra Firm Tofu - 14oz Nasoya Tofu, Organic, Extra Firm - 14 oz House Foods Tofu, Premium, Extra Firm - 16 oz
White Bread	Nature's Own 100% Whole Wheat Bread - 20 oz loaf Oroweat 100% Whole Wheat Bread - 24oz 100% Whole Wheat Bread 20oz - Market Pantry Great Value 100% Whole Wheat Bread, Round Top, 20 oz, Size: 12.1, 3.98, 4.21 Arnold Whole Grains Bread, 100% Whole Wheat - 1 lb

FIGURE 5: Sample inputs and their outputs for ingredients with commonly known healthier substitutes

FUTURE WORKS

This work could be improved by implementing Image Recognition Software (IR). As opposed to having users manually enter ingredients, an image recognition feature could scan recipes and report all ingredients with viable substitutions. Furthermore, image recognition software could identify similar recipes and suggest similar ingredient substitutions for them which would improve the speed and efficacy of this model.

In addition, the integration of a cross-checking feature in which the work checks multiple sources for the nutrition information of an ingredient would improve the accuracy of the work altogether. Similarly, allowing users to

check nutrition and allergen information for a product may be a next step. The cross-checking feature may also be useful in conjunction with the "allergen check" component to ensure that users are getting the most accurate allergen information.

ACKNOWLEDGEMENTS

The author wishes to thank Sara Metwalli for her guidance and aid in the research and production of the work, as well as for proofreading the paper.

REFERENCES

[1] *Public views about Americans' eating habits.* (2016, December 1). https://www.pewresearch.org/science/2016/12/01/public-views-about- americans-eating-habits/

[2] *MSU Food Literacy and Engagement Poll: Wave I – Food@MSU*. (n.d.). Retrieved August 28, 2020, from https://www.canr.msu.edu/news/msu-food-literacy-and-engagement-poll

[3] *5240107.pdf*. (n.d.). http://cs229.stanford.edu/proj2017/final-reports/5240107.pdf

[4] Rak, S. C. (2011). U.S. Patent Application No. 12/658,820.

[5] *Eurocode Main Food Groups: classification and policy, version 99/2*. (n.d.). Retrieved August 25, 2020, from http://www.ianunwin.demon.co.uk/eurocode/docmn/ec99/ecmgintr.htm

[6] EPA, U. S., & OW. (2016). *EPA-FDA Fish Advice: Technical Information*. https://www.epa.gov/fish-tech/epa-fda-fish-advice-technical-information