COSC – 6590, Application Development with Large Language Models

Pre-course Assignment

Submitted By: Biswas Poudel

1. In the file cnbc.html you will see a lot of places where you will find URLs of the type http:// … . If you want to extract only the URLs (anything starting with http), how would think about solving the problem. Either given algorithm or write Python code to do it and then count the number of such URLs. Come ready to discuss this on the first day of classes.

   Ans:

   Following algorithm can be used to extract only the URLs from the given html page.

i) Go to the html file of the webpage. On chrome, it is usually right click -> view page source (or Ctrl + U).

ii) In the body of the html file, look for the '<a>' tags. These tags contain the URL link under the attribute 'href'. So, the http:// link can be found in each line of the html file with '<a>' tag.

iii) After locating the '<a>', the 'href' attribute can be extracted to get the URLs.

iv) These URLs can be appended in a Python list. Thereby, at the end the list contains all the URLs in the webpage.

v) The length of the list returns the number of such URLs in the given html page.

2. Let us say you are setting up a grocery store online and you are required to virtually set up items for shopping online, so shoppers can search for them using names of the aisle categories. For instance, **eggs** may be mapped to the aisle **breakfast food** or **food**. The below file contains two columns: grocery items and aisle names. How would you map each grocery item to an aisle name? Think of an algorithm and see if you can write Python code for it.

   Ans:

   The algorithm for classifying grocery items is below:

i) Create a list of keywords for each of the aisle category. These keywords are the general items for the respective aisle category. For instance, the keywords for 'meat' could be 'chicken', 'beef', 'pork', 'lamb'. Similarly the keywords for rest of the category could be:

   Alcohol = wine, beer, vodka, whiskey, gin, rum, tequila, champagne, cider, ale, spirits, liquor

Baby = diapers, formula, baby food, pacifier, bib, baby wipes, bottle, infant, crib, stroller
Bakery = bread, baguette, croissant, cake, pastry, muffin, donut, pie, roll, bun, loaf

Baking = flour, sugar, baking powder, baking soda, yeast, cocoa, vanilla extract, frosting, sprinkles

Beverages = soda, juice, water, coffee, tea, lemonade, sports drink, energy drink, iced tea, soda pop

Candy = chocolate, gummy, lollipop, candy, toffee, caramel, licorice, sweets, hard candy, jelly bean, bar

Canned Goods = beans, soup, tomatoes, corn, peas, tuna, sardines, fruit cocktail, broth, stew

Cereal = oats, granola, muesli, cornflakes, bran, wheat, cereal bar, rice krispies, cheerios

Condiments = ketchup, mustard, mayonnaise, relish, barbecue sauce, hot sauce, soy sauce, salad dressing, vinegar

Dairy = milk, cheese, yogurt, butter, cream, sour cream, cottage cheese, kefir, buttermilk, lactose-free

Frozen = ice, frozen, pizza, frozen, ready-to-eat, dessert

Garden = seeds, fertilizer, soil, mulch, hose, planter, pruners, rake, gloves, potting mix

Household = detergent, soap, cleaner, trash bag, paper towel, toilet paper, sponge, disinfectant, mop, broom

Meat = beef, chicken, pork, lamb, turkey, bacon, sausage, ham, ground beef, steak, ribs

Misc = assorted, various, mixed, variety, combo, kit, set, miscellaneous, other, general

Organic = organic, natural, non-GMO, eco-friendly, pesticide-free, sustainable, grass-fed, free-range, hormone-free

Pasta & Grains = spaghetti, macaroni, noodles, rice, quinoa, barley, couscous, lasagna, penne, farro

Personal Care = shampoo, conditioner, toothpaste, deodorant, soap, lotion, razor, toothbrush, skincare, hygiene

Pet = dog food, cat food, pet toy, litter, pet bed, pet shampoo, pet treat, bird seed, aquarium, collar

Produce = apple, banana, carrot, lettuce, tomato, berry, cucumber, spinach, grape, potato, pepper, fruit, vegetable

Seafood = fish, shrimp, salmon, crab, lobster, tuna, scallop, oyster, prawn, clam, cod

Snacks = chips, popcorn, pretzel, cracker, granola bar, nuts, trail mix, snack mix, cheese puffs, jerky

Spices = salt, pepper, cinnamon, garlic powder, paprika, cumin, turmeric, chili powder, oregano, basil

ii) When classifying, the grocery items, look for any such keywords in the name of the items. If there is a match, the item belongs to the same category as the keywords.

iii) If there are partial matches with more than one category, a priority order can be implemented. For example, ‘ice cream’ could be matched with both ‘ice’ which is under ‘frozen’ category and ‘cream’ which is under ‘dairy’ category.
In this case, it can be placed in category where the first word of the item is matched. For ice cream, it’d be ‘frozen’.
Similarly, if it is ‘frozen vegetable’, it can be matched to the ‘frozen’ aisle although it will also be matched to ‘produce’ aisle with the keyword ‘vegetable’.

Other factors could also be considered for final decision when there are multiple matches.

iv) If no matches are found, they can be placed in ‘misc’ aisle. Or lookup the usage and description of the item on the internet and look for any keywords that appear in its description/usage.
After classifying the item, it can be included as one of the keywords for that category, for ease in the future provided any such items appear again.

This way, most of the items can be classified to a unique aisle. Although, some may still not be classified. In that case, a Machine Learning model can be trained to classify the items sold in major grocery stores. Although this would require a lot more work in acquiring the data, building model, and training it; however, it would complement the previous approach for more comprehensive and accurate classification. Then, these two methods can be used in combination.

P.S: ChatGPT was used to fetch all the keywords for each aisle category in the second problem.