Daisy Intelligence 2020 Hackathon Problem Statement

Background

In this era of digitization, storing, editing, indexing, and finding information in a digital document is much easier than spending hours scrolling through printed documents. Searching for something in a large non-digital document is not only time-consuming but is also prone to human error. Luckily for us, computers are getting better at doing these tasks every day. Today, almost everywhere, image recognition and Optical Character Recognition (OCR) are used to transform images into machine-readable information, thus optimising business processes by reducing tedious and mundane manual work.

In the following problem participants will work with scanned images of flyers for a grocery retailer. Participants will have to match the ad-block descriptions extracted from the flyers with the provided product dictionary and try to elicit additional meaningful information from all the flyers. They can test their solution against the example output file and will be scored based on the accuracy of the output fields generated from all the flyers.

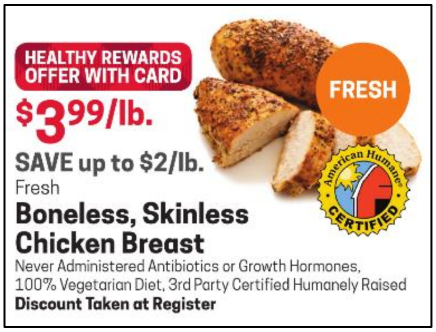
Problem Statement

Teams will be provided with a set of scanned images of flyers for a grocery retailer, a product dictionary that contains names of all the products, and a sample output file with some examples. Each flyer is for a weekly promotion and includes a start and an end date. Each flyer consists of multiple pages which in turn contain a varying number of ad-blocks. Each ad-block has a picture of the product, name and description of the product, price, unit of measurement (lb/kg or number of items), promotional discount, and additional tags such as organic, fresh, non-GMO, gluten-free, etc.

The product dictionary contains just the names of products. Participants must match ad-blocks using the descriptions on the flyer and the product dictionary and then extract the additional information for each ad-block. The output should be saved in a csv (comma delimited) format.

The description of the columns in the output file:

- **flyer_name** – Name of the image file from where the information is extracted without the file extension (e.g. input file – week1_page1.jpg, output value – week1_page1). This should be an exact match
- **product_name** – Name from the product dictionary that matches closest with the product name in the ad-block



HEALTHY REWARDS OFFER WITH CARD

$3.99/lb.

SAVE up to $2/lb.

Fresh

Boneless, Skinless Chicken Breast

Never Administered Antibiotics or Growth Hormones, 100% Vegetarian Diet, 3rd Party Certified Humanely Raised

Discount Taken at Register

e.g. ad-block product name – Boneless, Skinless Chicken Breast

product dictionary –

Boneless Chicken Breast
Chicken Breast
Chicken Thighs

output value - Boneless Chicken Breast

- **unit_promo_price** – Promotion price for each unit (e.g. ad-block - 2/$5, output value $2.5)
- **uom** – Unit of measurement of the product (e.g. lb, 1 Pint, 10 Pack, etc.)
- **least_unit_for_promo** – Least amount of the product the customer has to buy in order to use the promo (e.g. Save $3.5 on 2, output value – 2). Default value is 1.
- **save_per_unit** – amount of money saved per unit rounded off at 2 decimal places (e.g. - Save $3.5 on 2, output value – $1.75)
- **discount** – Discount on the original price rounded off to 2 decimal places (e.g. 2/$5, Save $6.98 on 2, output value = $6.98/$11.98 = 0.58)
- **organic** – 0 or 1 binary values indicating if the product is organic or not as described on the flyer. Default value is 0.

You will find some examples below in the Appendix section.

Objective

The participants are to submit an output csv file that contain the product names and relevant additional information as shown in the sample file and as instructed in the above section. Due to the complexity of the problem, it is unreasonable to have a perfect match between the ad-block descriptions and the product dictionary. As such, participants are recommended to use image recognition or OCR software to extract the text information from the flyers and perform a fuzzy match between the ad-block descriptions and product names from the Product dictionary. Participants are expected to submit the output for the flyers provided to them. The final score will depend on the number of products detected correctly and a weighted average of the accuracy of all the columns in the output file.

The flyer_name column values should exactly match the names of the files from which the data is extracted. This column will be used to match your output with the original data.

product_name column should exactly match how it appears on the flyer. If you don't find an exact match between the product name on the flyer and the product dictionary, use the name in the product dictionary that is a closest match. Refer to the example provided in the Problem statement section. product_name column carries the maximum weightage. If this value is wrong, the other columns will not be considered for your final score.

Presentation

The teams with the top scores will be selected to present their solution and how they came up with it. As such, each team should also provide a presentation of their solution and method, as well as provide any source code used to generate their solution. Teams are expected to be able to present their solution and method if they are chosen as one of the top teams. The presentation time should not exceed 10mins. The format of the presentation file will be a power point of 6-7 slides, and it must be included with the solution.

Submission

Teams will email their submission to hackathon@daisyintel.com This submission will include the names of the team members, an output csv file for all the ad-blocks, a power point presentation, and the source code for the algorithm used to generate the solution. You do not need to include any libraries used. Please zip together your files and name the zipped folder as solution_<team_name>.zip. Please make sure the name of the zipped folder does not contain any special characters. If your team name contains any special character replace it with alpha-numeric characters as you see fit. You will have until 10:00am to submit and you can only submit once. Files to submit:

1. output.csv - this file will contain the information for all the flyers. Please make sure the column names are exactly same as the sample_output.csv file.
2. presentation.ppt(x) – Your presentation on the solution which should typically contain 6-7 slides and should include name of the team members in the title slide.
3. Source code

File Format

The format of the product_dictionary , units_dictionary and sample_output file is csv. The product_dictionary contains the list of all products present in the flyers. The units_dictionary contains the list of possible (but not all) units of measurement present in the flyers. The sample output contains some examples and your output file should follow the exact same format. The first row of the output file contains the column headers. The subsequent rows consist of all the information of each ad-block.

The flyers are in JPEG format. Each flyer has multiple pages. The name of each JPG file suggests the week and the page number of a weekly flyer. For example – week_1_page_1.jpg means page 1 of the flyer for week 1.

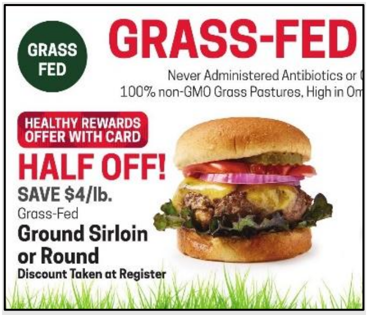
Materials

You can find all the material for this hackathon in the following google drive.

https://drive.google.com/open?id=1TZQPaLPI5rf9cDNU1im770CowLLqvs9u

It will contain this document, the flyer image files, and a sample output file.

Appendix



GRASS FED

GRASS-FED

Never Administered Antibiotics or

100% non-GMO Grass Pastures, High in Om

HEALTHY REWARDS OFFER WITH CARD

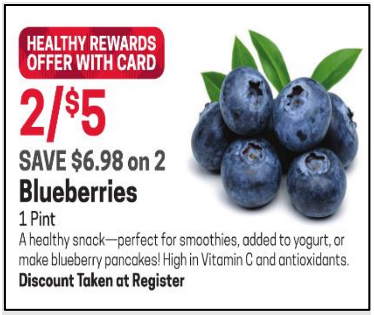
HALF OFF!

SAVE $4/lb.

Grass-Fed

Ground Sirloin or Round

Discount Taken at Register



HEALTHY REWARDS OFFER WITH CARD

2/$5

SAVE $6.98 on 2

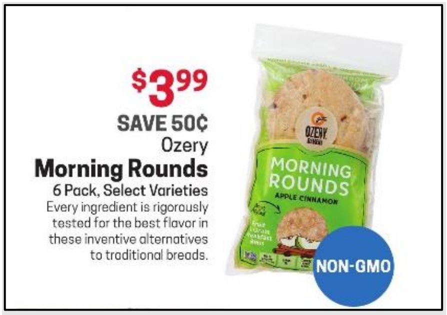
Blueberries

1 Pint

A healthy snack—perfect for smoothies, added to yogurt, or make blueberry pancakes! High in Vitamin C and antioxidants.

Discount Taken at Register

Ad-blocks from file week_1_page_1.jpg



$3.99

SAVE 50¢

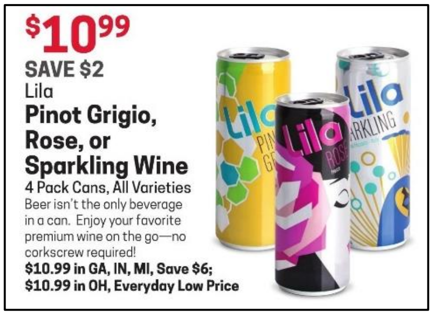
Ozery

Morning Rounds

6 Pack, Select Varieties

Every ingredient is rigorously tested for the best flavor in these inventive alternatives to traditional breads.

NON-GMO



$10.99

SAVE $2

Lila

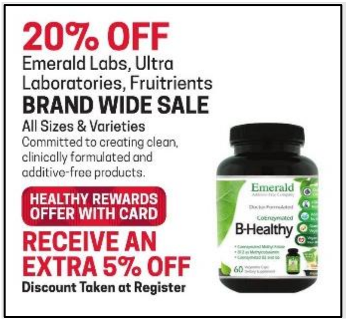
Pinot Grigio, Rose, or Sparkling Wine

4 Pack Cans, All Varieties

Beer isn't the only beverage in a can. Enjoy your favorite premium wine on the go—no corkscrew required!

$10.99 in GA, IN, MI, Save $6;
$10.99 in OH, Everyday Low Price

Ad-blocks from file week_1_page_2.jpg



20% OFF

Emerald Labs, Ultra Laboratories, Fruitrients

BRAND WIDE SALE

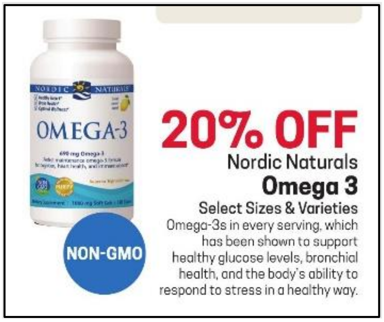
All Sizes & Varieties

Committed to creating clean, clinically formulated and additive-free products.

HEALTHY REWARDS OFFER WITH CARD

RECEIVE AN EXTRA 5% OFF

Discount Taken at Register



20% OFF

Nordic Naturals

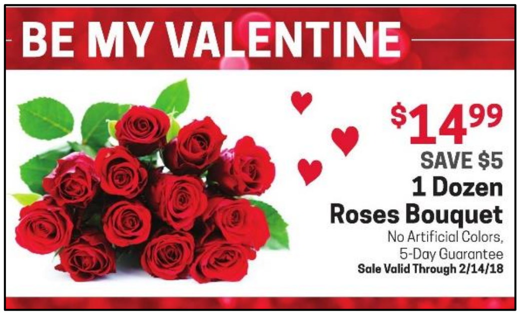
Omega 3

Select Sizes & Varieties

Omega-3s in every serving, which has been shown to support healthy glucose levels, bronchial health, and the body's ability to respond to stress in a healthy way.

NON-GMO

Ad-blocks from file week_1_page_3.jpg



BE MY VALENTINE

$14.99

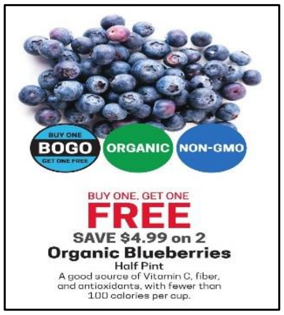
SAVE $5

1 Dozen

Roses Bouquet

No Artificial Colors,
5-Day Guarantee

Sale Valid Through 2/14/18



BUY ONE BOGO GET ONE FREE

ORGANIC

NON-GMO

BUY ONE, GET ONE

FREE

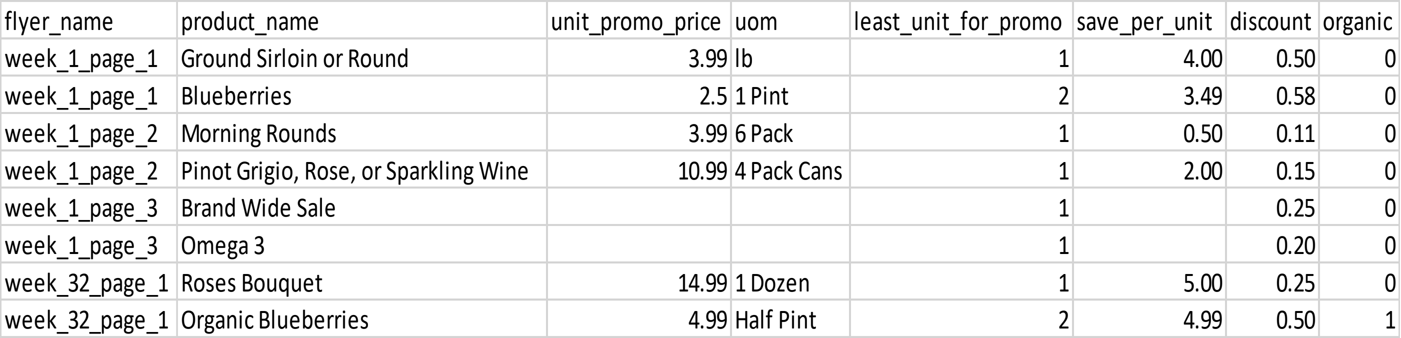
SAVE $4.99 on 2

Organic Blueberries

Half Pint

A good source of Vitamin C, fiber, and antioxidants, with fewer than 100 calories per cup.

Ad-blocks from file week_32_page_1.jpg



flyer_name	product_name	unit_promo_price	uom	least_unit_for_promo	save_per_unit	discount	organic
week_1_page_1	Ground Sirloin or Round	3.99	lb	1	4.00	0.50	0
week_1_page_1	Blueberries	2.5	1 Pint	2	3.49	0.58	0
week_1_page_2	Morning Rounds	3.99	6 Pack	1	0.50	0.11	0
week_1_page_2	Pinot Grigio, Rose, or Sparkling Wine	10.99	4 Pack Cans	1	2.00	0.15	0
week_1_page_3	Brand Wide Sale			1		0.25	0
week_1_page_3	Omega 3			1		0.20	0
week_32_page_1	Roses Bouquet	14.99	1 Dozen	1	5.00	0.25	0
week_32_page_1	Organic Blueberries	4.99	Half Pint	2	4.99	0.50	1

Sample Output