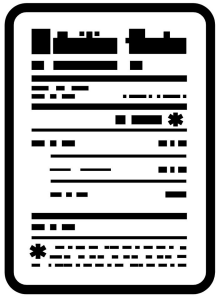
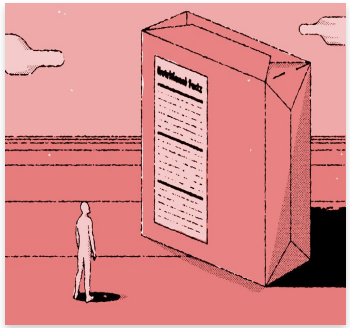


Datrition

Nutrition Label OCR

Omar Alharbi - Mansour Aljuaid -Sumaiah Alsadhan

TABLE OF CONTENTS



01

Problem Statement

Nutrition Facts Labels are not serving their purpose as it is hard to read and not all can interpret what it means ...

02

Project Idea

Utilize machine learning knowledge to enable consumers' make better decision while packaged-food shopping ...

03

Methodology

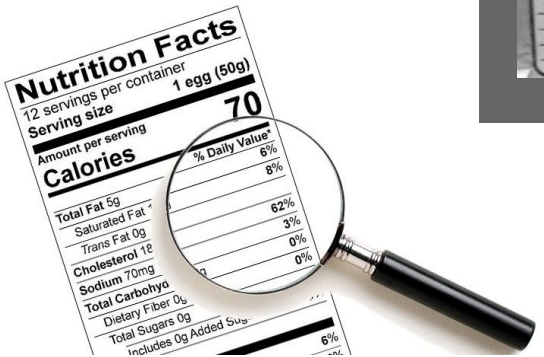
Used a combination of computer vision model, plus ocr model with focus on pre and post image processing ...

Nutritional Information Solutions

04

CONCLUSIONS

Built a dataframe and visualized the extracted data using the traffic sign system to aid decision making ...



Can you read it?
Do you understand it?
Are you able to decide?

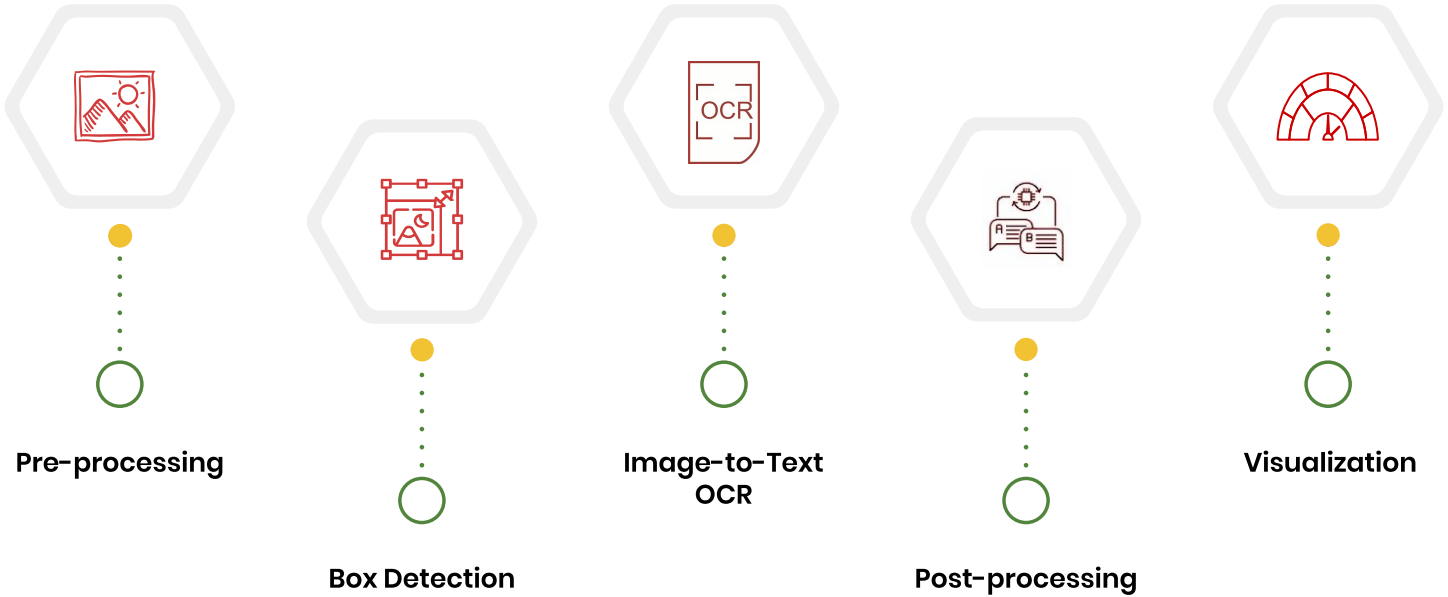
Everyday we come across nutrition facts labels but we often don't pay much attention to them. Sometimes, it is hard to read and other times even if it is large enough for you to read them. It is hard to tell if the product is the right one for you.

Project Idea

The idea is not complex. It is centered on the ability to detect nutrition label from an image. Then, extract the text into a dataframe. Finally, use visualization to aid decision making.

The idea was inspired from the SFDA datathon and our visualization is based on the traffic sign system established by many food and drug authorities.

Methodology



Using **OpenCV** and **Numpy**:

Thresholding

"**Thresholding** is a basic method for image segmentation, Before contour analysis an image has to be binary by converting it to a black and white image.

- ➔ **Outso** thresholding method avoids having to choose a threshold value and determines it automatically.
- ➔ **Adaptive Thresh Gaussian** the algorithm determines the threshold for a pixel based on a small region around it.

Box Detection (Find Contours)

"**Contours** are defined as the line joining all the points along the boundary of an image that are having the same intensity".

- ➔ Detecting boxes i.e. contours in a given image is applied through findContour () method in **OpenCV**.
- ➔ To Crop Nutrition Label the function will loop over all the contours,find the location of all the boxes at given dimensions and crop the part which has a rectangle.

Using **OCR engine:**

Pytesseract

Pytesseract is a python Optical Character Recognition (OCR) library. It is supported by google OCR engine (Tesseract), and is considered one of the most accurate open source OCR libraries.

Tesseract OCR

Image of OCR output

```
Nutrition Facts


Serving Size 1 (40g)
Anount Per Serving
Calories 120 Calories from Fat 15
4 Daily Values*
Total Fat 1.639 2.50%
Saturated Fat 0.24 1.2%
Trans Fat 0g Go.
Cholesterol 10mg 3.32%
Sodium 190mg 7.81%
Total Carbohydrate 25.129 8.37%
Dietary Fiber Og
Sugars 11.519
Protein 1.23

*Percent Daily Valued are basedon a 2,000 calorie diet
```

Nutrition Facts
Serving Size 1 (40g)

Amount Per Serving	
Calories 120	Calories from Fat 15
	% Daily Values*
Total Fat 1.63g	**2.50%**
Saturated Fat 0.24g	**1.2%**
Trans Fat 0g	**0%**
Cholesterol 10mg	**3.33%**
Sodium 190mg	**7.91%**
Total Carbohydrate 25.12g	**8.37%**
Dietary Fiber 0g	
Sugars 11.51g	
Protein 1.23g	

*Percent Daily Values are based on a 2,000 calorie diet.

Using **REGEX** and **Keyword Matching**:

Levenshtein Distance

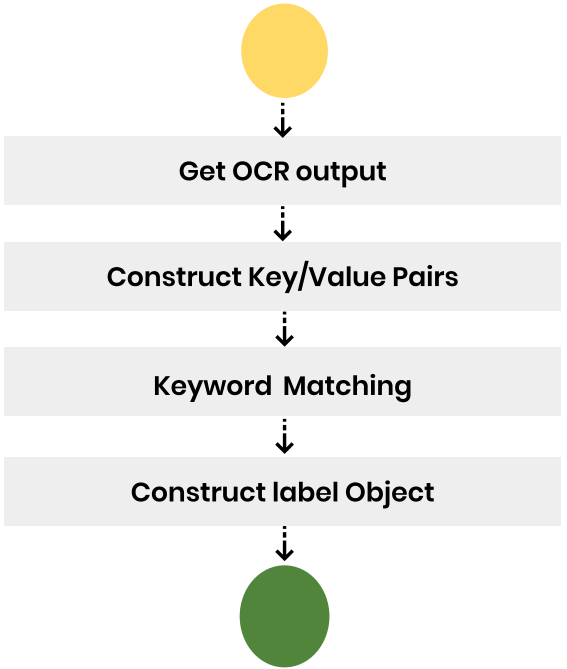
The Levenshtein (edit) distance is a metric that reflects how closely a given string of characters matches another given string.

The first step of preprocessing is to read through the OCR output and decide which information is important and which information to be ignored..

The next step uses the identified important lines of information and create a name (key) and number (value) from each line. The use of an algorithm that apply levenshtein distance to match the text to the given list of nutritions.

The outcome is a dictionary object of keys and values.

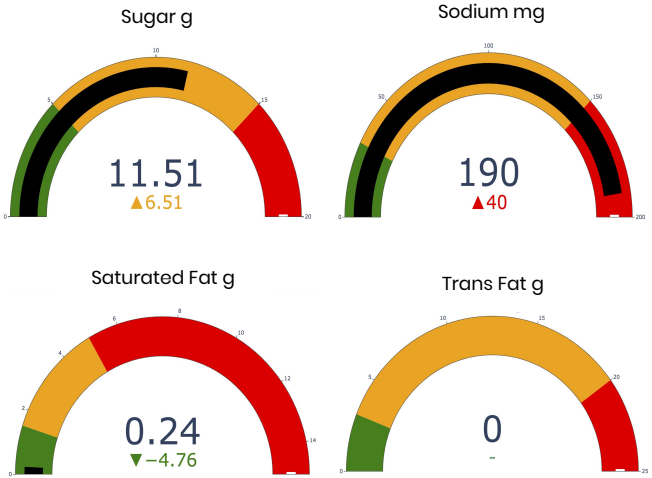
Activity Diagram of post process



Nutrition Gauge

Nutrition Label Example

	Sugara	Sodium	Saturated Fat	Trans Fat
CAKERS Cupcake	11.51 g	190 mg	0.24 g	0 g



UNDERSTANDING THE TRAFFIC LIGHT SYSTEM

	Sugars	Fat	Saturates	Salt
What is HIGH per 100g?	Over 15g	Over 20g	Over 5g	Over 1.5g
What is MEDIUM per 100g?	Between 5g and 15g	Between 3g and 20g	Between 1.5g and 5g	Between 0.3g and 1.5g
What is LOW per 100g?	5g and below	3g and below	1.5g and below	0.3g and below

Source: Food Standards Agency

Challenges

Model Generalization	This is still a research field. How to apply the concept not on one image but a folder containing hundreds..
Computational Power	To process many images; it requires the utilization of GPUs and not CPUs
Improve Accuracy	Extensive work need to be done at the pre and post image processing stages.
Experimentation	This is just the start. We would like to examine other approaches that we opted out off to remain focus..

Opportunities

Developments	Link the medical record of the users to raise awareness in making better healthy choices..
Application	An application that serves healthcare, retailers, pharmacies and consumers
Platform	A platform that can be used by government bodies for a diverse range of purposes e.g. product registration, tax levy.

Thank you for your attention!

Any Questions?