A black background with grey leaves

Description automatically generated

Storing Extracted Nutritional Data in Tables

Week 5

A grey logo on a black background

Description automatically generated

February 26, 2024

Divyanshu

Contents

[Objective 3](#_Toc160017176)

[Implementation Overview 3](#_Toc160017177)

[AWS Textract Configuration 3](#_Toc160017178)

[Image Processing 3](#_Toc160017179)

[Accuracy Calculation 3](#_Toc160017180)

[Nutritional Information Extraction 3](#_Toc160017181)

[Results and Output 3](#_Toc160017182)

[Image Analysis 3](#_Toc160017183)

[Image 1 – *21040\_Dairygold Original - 63\_ Fat Blended Spread\_back* 3](#_Toc160017184)

[Image 2 – *21046\_Centra Light Irish Milk\_Nutritional Info* 4](#_Toc160017185)

[Image 3 – *21080\_Hellmann\_s Light Mayonnaise\_Nutritional Information\_600g\_Glass \_ Plastic* 5](#_Toc160017186)

[Image 4 – *21124\_Avonmore Low Fat Super Milk\_Back* 6](#_Toc160017187)

[Image 5 – *21260\_Weetabix Original\_Nutritional Info* 7](#_Toc160017188)

[Image 6 – *21567\_Glenisk Organic Kids Bio No Added Sugar Vanilla Yogurt\_Back (1)* 9](#_Toc160017189)

[Image 7 – *21725\_Lucozade Sport Orange\_Nutritional Information\_750ml\_Plastic* 9](#_Toc160017190)

[Image 8 – *22548\_Tesco Bran Flakes\_750g\_Nutritional Information\_Paper \_ Board* 10](#_Toc160017191)

[Image 9 – *Kellogg\_s Bran Flakes\_Nutritional Info* 12](#_Toc160017192)

# Objective

The objective of this report is to extract text from food nutrition labels using the AWS Textract tool, store the information in a CSV file, and evaluate the accuracy of the extracted text compared to ground truth data using the Levenshtein distance algorithm. Additionally, The report also involves highlighting errors in the extracted text.

# Implementation Overview

## AWS Textract Configuration

The AWS Textract client is configured with appropriate AWS credentials, including the access key ID and secret access key. The Textract client is set to operate in the 'us-east-1' region.

## Image Processing

Images are read as byte arrays using the read\_image\_as\_bytearray function. The Textract tool is used to detect document text in the images, and the resulting text and confidence scores are extracted.

## Accuracy Calculation

The Levenshtein distance algorithm is employed to calculate the accuracy of the extracted text compared to ground truth data. The calculate\_accuracy\_and\_errors function is used for this purpose.

## Nutritional Information Extraction

The extract\_nutritional\_info function is responsible for extracting nutritional information from the text. It uses predefined nutrient keywords and searches for them in the text to extract corresponding values and confidence scores.

## Results and Output

The extracted text, Levenshtein accuracy, and errors are printed for each image. Additionally, the nutritional information is extracted and stored in a CSV file for further analysis.

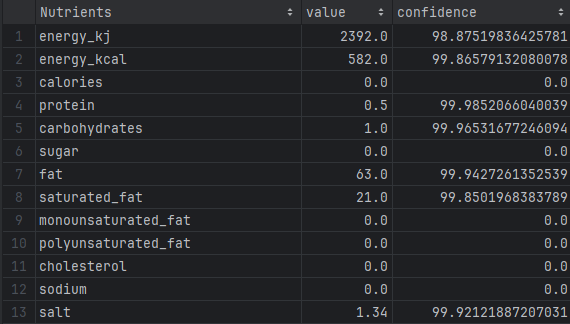
## Image Analysis

### Image 1 – *21040\_Dairygold Original - 63\_ Fat Blended Spread\_back*

A label with information on it

Description automatically generated

#### Output:



### Image 2 – *21046\_Centra Light Irish Milk\_Nutritional Info*

A green label with white text

Description automatically generated

#### Output:

A screenshot of a computer

Description automatically generated

### Image 3 – *21080\_Hellmann\_s Light Mayonnaise\_Nutritional Information\_600g\_Glass \_ Plastic*

A blue label with black text

Description automatically generated

#### Output: A screenshot of a computer Description automatically generated

### Image 4 – *21124\_Avonmore Low Fat Super Milk\_Back*

A red and white label with white text

Description automatically generated

* Salt was read as **0.16** instead of **0.1g.**
* Energy wasn’t extracted.

#### Output:

A screenshot of a computer program

Description automatically generated

### Image 5 – *21260\_Weetabix Original\_Nutritional Info*

A yellow label with black text

Description automatically generated

* Energy in kJ calculated as **1514.**
* Carbohydrates as **26** instead of **69.**
* Fibre extracted as **12** instead of **10** which is protein’s value.
* Niacin extracted as **0.8** instead of **14mg.**
* Folic acid extracted as **0.6** instead of **170µg.**
* Iron extracted as **100** instead of **12mg.**

#### Output:

A screen shot of a computer

Description automatically generated

### Image 6 – *21567\_Glenisk Organic Kids Bio No Added Sugar Vanilla Yogurt\_Back (1)*

A label with white text and white text

Description automatically generated

#### Output:

A screenshot of a computer

Description automatically generated

### Image 7 – *21725\_Lucozade Sport Orange\_Nutritional Information\_750ml\_Plastic*

A blue label with white text

Description automatically generated

* Nothing was extracted.

#### Output:

A screenshot of a computer

Description automatically generated

### Image 8 – *22548\_Tesco Bran Flakes\_750g\_Nutritional Information\_Paper \_ Board*

A blue label with white text

Description automatically generated

* Energy kcal calculated as **358.75.**
* Niacin wasn’t extracted.
* Vitamin b6 wasn’t extracted.
* Folic acid wasn’t extracted.
* Vitamin b12 wasn’t extracted.

#### Output:

A screen shot of a computer

Description automatically generated

### Image 9 – *Kellogg\_s Bran Flakes\_Nutritional Info*

A blue label with black text

Description automatically generated

* Energy kcal calculated as **362.57.**

#### Output:

**A screenshot of a computer

Description automatically generated**