**Groceries Masters: Spread Sheet Generation (OCR)**

**Things to put in consideration:**

Skew correct

**How it works.**

There are two images of the same picture that will go through the OCR (normal and binary (black and white))

**Binary vs Normal**

The binary has accurate numbers, while the format is more off. While the normal has a more accurate layout, while the numbers are more off.

Start by storing the prices gained from the parsed binary image in to a pricelist. This will now become your bank of price lists.

**Check for a possible total cost?**

Second, do the same with the normal image, however this time, tuple the price with the name of the food. The name can be extracted using this algorithm.

Starting String: (E 24311 R. MUFFIN 15.90)

Remove all spaces from the line (E24311R.MUFFIN15.90)

Extract the price, from each line as done before and store it (Extracted: 15.90)

If price is extractable, find the index of the first character of price string. (index 14)

Remove everything after that index on the line. (E24311R.MUFFIN)

If first character is E, remove it. (24311R.MUFFIN)

Remove all numbers and punctuation/special character (RMUFFIN)

If the length of above value is <2 then most likely it’s a coupon (keep it in storage, throw out for now?)

Tuple RMUFFIN with 15.90 to get (RMUFFIN, 15.90)

\*\*Maybe should keep track of current position of the tuples, so that they can be ordered in the future and correspond with the receipt.

By now you have a pricelist from the binary image and a list of tuples from the normal image, start by comparing, the values of the pricelist with the prices on the tuples. Split the list of tuples into two lists: Ones with prices in the pricelist, and Ones that don’t. Pop out, values from the pricelist if they are in the tuples list.

For the remaining tuples assess the similarity and choose closest looking price from price list.