Product Categorization

NLP based product categorization is an essential step for structuring tax receipt information and utilizing huge potential hidden in that data. In order to achieve that several steps were implemented:
Important: *the variety of products in final consumption sales documents (HDM) is huge. Only in one supermarket chain the unique products (product lines) sold over three month period were more than 50,000. Overall, to ensure maximum benefit it was decided to start product categorization with the largest FMCG store chains (supermarkets). According to our estimates these supermarkets cover 30 % of overall final consumption in the country, and more than 80% of essential products consumption.*
Steps for product categorization already implemented:
Scraping
Getting alternative data with scraping based on selenium for having reference dataset from two main supermarket chains- sas and city. However, as their structure is a bit different we created a new column, *Final Category*, so that the matching can be consistent.
In addition, despite product level information we scraped the brand names in Armenian, Russian, and English languages for separately identifying brand names from the receipts alongside with the production country.
Data cleansing
We implement basic cleansing on Armenian text from removing noise, such as removing specific words, product codes, punctuation and special characters, values enclosed in parenthesis and mapping unit measures for consistency (both for English and Armenian), such as g, gr, or grams to gram, etc.
Translation
Translating Armenian text into English for having larger opportunities to work with state-of-the art NLP algorithms. We implement this step by checking two translation algorithms in case one of them is not working, the code will use the second one.
Overall, this is quite a time consuming process even with the parallel processing, on average translating 180,000 unique product names takes approximately 3 hours.
Data cleansing
Cleansing translated text for removing additional noise by removing digits, re-mapping unit measures, removing additional spaces and replacing some stand alone letters, such as g-gram, l-litr.
Main subject extraction
In order to get structured information from tax receipts it was decided to implement transformer based sentence similarity embeddings based on cosine similarity. Several different combinations of matching approaches were tested and the final pipeline consists of three main steps.

1. Matching cleaned-translated product name to cleaned products from supermarkets
2. Matching cleaned-translated product name to final categories
3. Comparing which one of the two approaches have higher cosine similarity score and for the final group choosing that one.

This approach not only enables a more general method for main topic identification but also enables filtering based on the matching score; higher the score more certain is the model that products are similar.
Brand name and unit measure extraction
Although product is one of the main subjects in the text unit measure and brand and product country is very crucial, as a result this was separately handled.
Unit Measure Extraction

- We have obtained a list of all the possible unit measure a product can have
- Using python regular expressions we are retrieving the unit measure and its value from the product description (note that a product might have more that 1 unit measures, example “sausage 10pcs 20gr” we are extracting up to 4 unit measures from the product description (more doesn't make sense and is not present in the data)
- Besides the unit measure we also extract the value itself for the example mentioned above the output will be [(10, pcs), (20, gr)]

Brand name and product country extraction
Brand name matching is done by the following steps

- Extracting brand names enclosed within special characters (examples: [“brand_name”, ‘brand_name’, \ <brand_name>, etc.])
- Exact match for armenian brands list (the brand list is obtained by scraping supermarket data (city, sas))
- Exact match for english brands list
- Match armenian brands to English brands to obtain final brand name in english
- Merge the final brand with the Origin Country data obtained from supermarket data scraping

Future Actions

- More targeted cleansing for removing adjectives, for example apple juice or apple cookies sometimes are matched with apple instead of juice or cookie.
- Manually double check matching dictionaries and remove misleading candidates and checking final categories
- Separate approach for each category, such as rule based for fruits and vegetables or some price filter for higher categories,