UNIT MEASURE EXTRACTION

One of the initial steps of processing textual data from HDM checks is extraction of unit measures and their respective value from the textual data. For this purpose unit measure extraction and unit measure standardisation methods were implemented.

Unit Measure Extraction

Unit measure extraction consists of multiple parts. First the textual data is cleaned and transformed to make the extraction process easy. Then the extraction happens with two steps.

1. Firstly we extract unit measures using Qunatulum3 library, a tool for extraction of unit measures. While Quantulum is well suited for extracting unit measure and numerical value pairs, it is not effective for instances where numerical value is not specified.
2. Therefore for the next step we extract unit measures and their corresponding values using regular expressions. Here we concentrate on extracting Unit Measures even if the numerical values are not attached.
3. As a complementary step we try to extract the size of the product if it is mentioned (this is a place for future improvement). Example: paper towel 20 x 30 mm Extraction→ 20 x 30 mm
4. After completing the extraction steps we sort the extracted units by importance, assuming that weight and volume indicating unit measures are the most important. This way at the end of this process we get multiple possible options for unit measure which are sorted by the importance.

Unit Measure Standardisation

After Extracting the possible unit measures we need to find out which of the extracted measures are the most reliable to be labelled as the true unit measure of the product, and the values need to be standardised. Here the algorithm passes through multiple steps to obtain the most reliable unit measure of the product.

1. First we define the rules by which each unit should be standardised, bringing every product to the same unit of measure in their category (mass - gram, volume ml etc.)
2. After defining the rules the process of finalisation starts. At first we check the first extracted unit measure which is already sorted as the most relevant in the previous part of the program. Check if that unit measure passes the rule of standardisation, and if it does we assign that unit measure to the product.
3. If the measure is rejected by rule, the algorithm falls back to the original unit measure column from the original data. If the original unit measure is in the accepted format (indicating mass or volume) it is accepted as the final unit measure.
4. If the desired unit measure is not found from the previous steps, the algorithm starts to search a unit measure that passes the rules, in the other extracted measures. Algorithm stops if it finds a unit measure and numerical value that satisfy the predefined rules.
5. After passing through checks if the unit measure is still not found the algorithm does one more fallback and checks whether the first Unit Measure and the numerical value can be accepted in case of some transformations. If the condition is satisfied, the algorithm completes the transformation and assigns Unit measure and the values.

6. If the unit measure is not found in all of the previous steps the program does not return any unit measure for that specific product.
7. As an additional step, the program adds an additional information column which returns the information about product size extracted in the first part if it exists, and adds the percentage information extracted from data.