Task 1c

Vinay Pinjani | 1151832

Scoring Method: How the product comparison works

This part of the task makes use of the "textdistance" library. The threshold set for the minimum normalized similarity was 0.75 and the scoring function of choice was the "Cosine Similarity".

Analyzing the records in the "abt_small" dataset, it was realized that the last element of the name string contained a unique product identifier and the first element of the name string contained its manufacturer. The comparisons between records were based on these elements.

The programs run a nested for loops to compare all possible records in the 2 data frames. Using regex patterns, all punctuations and cases were ignored. If the unique identifier from "abt_small" record was was also present in the "buy_small" records, then the two record were matched. This was most records that exactly matched were found. If a match was not found, then, for the elements with the same manufacturers, the unique identifier was compared with the "buy_small" record, and the maximum of the similarities produced was taken. If this similarity was greater than the threshold set, then the records were matched. This way the records where the unique identifiers were slightly different were also matched.

Evaluation of the overall performance of the product comparison

Using the described method, we get a recall value of 0.84 and a precision value of 0.87, specific to the given dataset. The precision may be due to the threshold value of 0.75 which may allow false positives. This can be improved by using a higher threshold value. The recall value can be improved by generating more matches, this may be done by ngrams to compare the unique identifier or by reducing the threshold further, this will allow lesser false negatives in our data.

An approach that may improve our results can be do individually compare the names and description (if available) and use the maximum of the 2 similarity values obtained. This products with more common attributes will be matched.

Blocking method: How the blocking method works

This part of the task makes use of the "manufacturer" attribute in the "buy.csv" dataset. If the "manufacturer" attribute is null, the row is given a manufacturer by using the first element of the name of the records. All elements of the manufacturer column are converted to lower case strings and in case where manufacturer name as multiple words, only the first word is taken, this ensures that all manufactures are in the same format.

Blocks are formed of all the products containing the unique manufacturers using the “groupby” method and a list of these unique manufacturers is stored. For each of the record in “abt.csv”, it is checked if the name contains any of the manufacturer for the list using regex patterns (removing punctuations and ignoring case) and if found they are assigned to that manufacturers block. The particular manufacturer is used as key blocks in both data sets.

Evaluation of the blocking method

Using the described method, we obtain Pair Completeness value of 0.94 and Reduction Ratio of 0.95, for the particular dataset. The low value of Pair Completeness may be due to losing the records that do not contain the manufacturer name, a way to overcome this is by using a secondary attribute such as “price” to form a block.

The time complexity of the blocking method linear, as only name strings of “abt.csv” are being compared with blocks formed.