README - Product Duplicate Detection by Locality Sensitive Hashing

Computer Science for Business Analytics

Author:
Sterre Spaanderman

Student number:
510941

This project determines which products, found on different websites, are duplicates. In particular, we aim to decrease the computational time by implementing LSH. Then, we will apply MSM to determine the dissimilarities between potential duplicates. Then we will cluster the products that are fairly similar. Finally, we implement multiple measures in order to determine the performance of the model.

The code is structured in the following way. The first part of the code inserts and cleans the data, as mentioned in the paper. Then, the LSH is implemented; we created permutations, and found what pairs of products are potential duplicates. Next, the dissimilarity matrix is created, for which the non-potential duplicates obtain a value of infinity. Then, the MSM is implemented for all pairs of products in the similarity matrix that have not obtained a value of infinity. The dissimilarity matrix is then formatted properly for clustering. Subsequently, the single linkage hierarchical clustering algorithm is implemented, for which we obtain the duplicates. Finally, the performance measures are implemented, and the plots are obtained.

For the remainder of the code, we will implement the same algorithm, but now for the SD.

The code is used in the following way. One can perform the code on a dataset that has the same layout as the one obtained from [1].

References

[1] Fransincar. Dataset webshop television desciptions. 2022.