This project is about the methods locality-sensitive hashing and duplicate detection. The dataset contains a product description of 1624 televisions and the goal is to find the duplicates from this dataset. The first part of the code imports the data and performs data cleaning to make all the descriptions consistent. Then the model\_ID and webshops are extracted from the data followed by all the functions used in the model. The description of each function are described in the following table.

|  |  |
| --- | --- |
| Function | Description |
| 1 | Splits all the titles into the substrings |
| 2 | Creates model words from substrings |
| 3 | Creates binary matrix |
| 4 | Creates signature matrix |
| 5 | Creates number of bands and rows to consider |
| 6 | Creates the hash-buckets |
| 7 | Creates the candidate pairs and candidate matrix |
| 8 | Computes Jaccard (dis)similarity and creates dissimilarity matrix |
| 9 | Clustering of candidate pairs |
| 10 | Evaluation of LSH |
| 11 | Evaluation of clustering |
| 12 | Bootstrapping |

Then the code continues with extracting the titles from the model and the distance threshold for clustering is determined. Finally the model is evaluated using bootstrapped data and the graphs are created.

To use this code one should first run all the parts above the main section, so run the data import and data cleaning part and all the functions. Then one can run the main part in order to see the results.