Report

Duplicate Detection

A duplicate instance is an exact replica of another instance in terms of specification. Our task was to identify duplicates in the given data set.

The method of cosine similarity is used to identify duplicates in the dataset. The cosine similarity is performed on the ‘title’ and ‘detailedSpecStr’ attributes as these identify an instance uniquely. Cosine similarity seems to be most suitable for this problem as here only the keywords in the title and description matter not the relation among them.

Since the dataset contains approx. 3.5 lacs instances, due to which to detect duplicates each instance has to be compared to every other instance. But this can be avoided and the total complexity can be reduced. This is done by splitting the data frame according to different ‘sleeve’ types(since clothes with different sleeves won’t be duplicates). Similarly, the resultant data frame is again split by different ‘neck’ types and then size(size attribute can also be considered for duplicates). This will reduce the total number of comparisons since the comparisons will be done only with instances with the same ‘sleeve’, ‘neck’ and ‘size’ attribute.

The cosine similarity is performed by converting the title, detailedSpecStr into tokens and generating thier TF IDF matrix and calculating their cosine product. In this solution, exact duplicates are considered so only duplicates with 100% match are considered. But near-duplicates can also be detected by defining a threshold.

Images can also be used for duplicate detection but on a large dataset like these would require good computing power.

NOTE: The duplicate10000.txt contains the result of the program on first 10000 instances(due to limited computing power).