Infilect Duplicate Detection Assignment WriteUp

I have read online that duplicate ecommerce product pages can be created using same img urls and description and even same title for different sizes of shirts.

That case can be handled using groupby of description,sellerName and title.

Other Difficult form of Duplicate content is same image with different colors are uploaded as different web pages .

Product Descriptions are same for different tops and yet different for same pattern /design tops on urls.

So description is not a great feature to discriminate duplicate products

Some sellers uploaded same description for different products altogether .

So groupingby description wont work in this case of duplicate detection.

I haven't had chance to explore all the columns thoroughly but initially keySpecsStr showed promise for clustering of similar pattern tops.

I removed columns where most rows are null i.e sizeunit ,DeliveryTime etc.

I also observed that tops 'title' column names are formed as [product_Brand]+[sleeve type]+[fabric_type]+[colour]

And most of the features from keySpecsStr like sleeve ,neck ,have choosen to form title of tops in duplicate pages.

So i choose productBrand,sellerName,keySpecsStr as main columns to groupby .

Then i observed that it almost segregated similar patterned tops together .

After looking at grouped img urls i came know that it is important to extract pattern on tops of models to compare different images.

I used local_binary_patterns algorithm to obtain histograms and ranked similarity between grouped/indexed images.

I observed that local_binary_patterns worked when image colours didn't variated much.