1 Feature Selection

A few columns have been selected for identifying duplicate records.
There are multiple sellers selling the same item. Different sellers will sell the same product at different prices. We will not consider the price of the product as a parameter for duplicate detection.
For any two products if the similarity score is greater than certain threshold then we consider one of them as duplicate entry.

"inStock":- One item can be sold by two seller. One seller might have the item present in his/her inventory the other might not have it. We would like to keep one of the items. Hence ignoring the feature "inStock".
One product can be sold by different sellers. In this case users/customers need to have a choice to select among the sellers. However, if any two items have the same "sellerName" along with same value for some other attributes then one of them is a duplicate entry.Thus including "sellerName" in the list.
A few features that have been considered are as follows:-

- title
- description
- productFamily
- productBrand
- keySpecsStr
- detailedSpecsStr
- Size
- categories
- productURL
- sleeve
- neck
- color

2 Using "Spacy" for Duplicate Detection

Spacy internally uses Cosine Similarity in order to calculate the similarity between two vectors. We need to convert each product into a vector. This is achieved by the library "word2vec".

The cosine similarity between two vectors is expressed as below:-

$$CosineSimilarity = (\vec{u}.\vec{v})/(|\vec{u}| * |\vec{v}|) \quad (1)$$

$\vec{u}$ and $\vec{v}$ represent the two items(/products).

3 Result

The result is stores in two files "Similar" and "Different". All the products having the similarity ¿0.95 are considered as duplicates and are placed in the file Similar. The rest are placed in the file Different.