Duplicate Detection of Product Listings

Task :- We want to find the duplicate entries of Jeans in the dataset. Since the output labels are not given i.e. we donot know which jean is similar to which jean. This is a problem of Unsupervised Learning.

Approach:-

1. Select only the category that contains Jeans.
2. Now, Remove the columns which are unnecessary for Jeans.
3. Remove the Columns that are NAN.
4. Do, One hot Encoding for categorical varibles.
5. Do Dimensionality Reduction
6. Split data in Train and Test
7. Apply Machine Learning on Train Dataset
8. Evalution using Silhoute Coefficient
9. Comparison between the Approaches

In between the approaches I have asked some question and given answer to them because those question and answer display the way of tackling the problem.

Detailed Approach :-

Columns in Dataset :-

```
'productId', 'title', 'description','imageUrlStr', 'mrp','sellingPrice', '
specialPrice', 'productUrl', 'categories','productBrand', 'productFamily',
'inStock', 'codAvailable', 'offers','discount', 'shippingCharges', 'delive
ryTime', 'size', 'color','sizeUnit', 'storage', 'displaySize', 'keySpecsSt
r','detailedSpecsStr','specificationList', 'sellerName', 'sellerAverageRat
ing','sellerNoOfRatings', 'sellerNoOfReviews','sleeve', 'neck', 'idealFor'
```

First I selected only the categories that contain Jeans.

```
'Apparels>Kids>Boys>Jeans',
'Apparels>Kids>Girls>Jeans',
'Apparels>Men>Jeans',
'Apparels>Women>Fusion Wear>Jeans & Shorts>Jeans',
'Apparels>Women>Maternity Wear>Jeans & Shorts>Jeans',
'Apparels>Women>Western Wear>Jeans & Shorts>Jeans'
```

Now, I since we know that jeans donot have sleeve, neck. I removed these columns.

Now, I saw the columns that are null. I found that these columns are null :-

'displaySize','specificationList','sleeve', 'sizeUnit','storage', 'offers','discount','shippingCharges', 'deliveryTime',

The following feature has all values as True :- 'codAvailable'

'Size' was an important feature but due to time constraint I had to remove this feature.

Further Exploring the dataset. I found out that 'detailedSpecsStr' is exactly same as keySpecsStr. So, I removed 'detailedSpecsStr'.

'ProductUrl' will be different for all the products weather they are same or not. So, this feature plays no role in identifying duplicate. So, I removed this feature.

Saving the DataFrame :-

Now, I was left with only 17 columns and 68559 rows.
I saved this dataframe in a file because I don't want to load the 5 gb file every time when I run the code .

One Hot Encoding :-

Since, Sklearn takes all feature as integers. So, I need to do **One Hot Encoding** of the features.
Features that need to be one hot encoded are :-
'Color', 'keySpecsStr','sellerName','title','imageUrlStr','categories','productBrand'.

I applied one Hot Encoding on :-

1. sellerName
2. Categories
3. ProductBrand

I did not do One Hot Encoding on Color, KeySpecsStr, imageUrlStr,title because they were creating mem ory problems in my Laptop.

Question Why I selected the above three features only for one hot Encoding?
Ans. Different products have same color, so choosing it was an option when I had memory. imageUrlStr i s only 13 in number , this tells that there are various sellers using same photo so, this does not lead to good feature.
Title and KeySpecsStr , I would have used them in other models if I have more time, in place of ProductB rand.
Categories : I used it because of its low dimensionality, It would be much memory efficient.

Now, My data is of shape:-

```
Rows :68559
Columns: 1029
```

Now, I would drop Color, KeySpecsStr, title, imageUrlStr. Because sklearn takes only integer as inputs.

Dimensionality Reduction:-

Question:- Can I use PCA here to reduce the dimension of the dataset.
Answer :- No, we can't use PCA here, because PCA donot work well on categorical variable because cate gorical varibles are non Guassian.

Question:- Then What can we do to reduce the dimesion of the dataset?
Answer We could use Multiple Factor Analysis.

Now, After all the pre-processing is done. Its time to apply Machine Learning.

Since, there is no label in the problem , hence it is unsupervised learning. We have to detect the Jeans which are similar.

K nearest neighbor :- I assumed that there are 2000 different kinds of Jeans in dataset of 68559. Now, I applied K nearest neighbor to find the jeans that are similar.

Now, I split dataset in 2 sets, training set and testing set.
Train set size = 40000
Test set size = 24000

Now, I will apply K mean clustering on train set size.

Evaluation Metric:-

Silhoutte Coefficient :- If the ground truth labels are not known, evaluation must be performed using t he model itself. The Silhouette Coefficient is an example of such an evaluation, where a higher Silhouett e Coefficient score relates to a model with better defined clusters. (Source :- Skleran library)

Neural Network :-

If I had time, then I would have used neural network.

Neural Network Vs K nearest neighbor

Neural network would have worked better because in K nearest neighbor we are finding clusters, so we are getting the group of objects that are similar. For example :- If Jean1, Jean2, Jean3 are identified in sa me class then we are able to find that these 3 are similar. We are also able to tell that Jean3 belong to th is class with 80% probability but we are unable to say what is the relation between the similarity betwee n Jean1 and Jean3.
Whereas in neural network we are able to tell this.