This problem statement deals with the problem of duplicate product listings in an e-commerce inventory.

**Tasks:**

Download the data file and perform all your operations on **subcategory:**

• As part of this problem statement, we need to come up with a definition of a duplicate as well. Feel free to search on the web how others have defined and solved **duplicate detection problem**.

• Note that, the dataset has its peculiarities (e.g., some fields/columns-values may repeat across rows.) Also, note that the dataset is quite rich and you may need to do substantial scripting to parse the data.

• As your process the dataset, you will realize that there is at least one data type that you can use very effectively to define and solve duplicate detection problem. If you identify the most amenable field(s) to do duplicate detection, you are onto a good start. Hint: Use images.

• You may want to perform the duplicate detection task using machine learning techniques (for eg. word2vec, CNNs, RF) for feature vector extraction. Note that, you may or may not be required to do custom training.

• If you use more than one approach to solve the problem, please do a quantitative comparative study as to why one approach works better than the other(s). If not i.e. if you have used only one technique, convince us as to why this technique would perform better than the rest.

• Think about building a function that takes two products/rows as input and produces output as yes or no (be careful with what is a product, please look at the dataset carefully, there are some obvious data pre-processing steps you will have to execute to answer 'what is a product'). How many comparisons you will have to do for each row? If there are N rows, how many such comparisons you will have to do to find all duplicates?

• If you feel that there is no labelled data, don't spend time in creating any labelled data. The dataset is self-contained and the assignment does not require to put efforts in creating in any dataset.

• Once you figure out how to identify duplicates (or approximate duplicates), take a look at a sample JSON file(sample.txt) which is a dictionary with a product id as key and list of tuple(s) of duplicate product id.

**Remember:**

• Pay keen attention to data before designing pipeline.

• Using a ML technique does not necessarily mean training a new model. Please feel free to use any pre-trained model for any task

• You may want to use text as well as images in addition to other fields

**Code of Honour:**

You may consult any papers, books, online references, or publicly available implementations for ideas and code that you may want to incorporate into your strategy or algorithm, as long as you clearly cite your sources in your code and your writeup.

**P.S. Machine Learning is very subjective. Whatever works and gives lowest error rate is the best approach and it varies from dataset to dataset. Try your best to achieve the best accuracy.**