NLP-Task

1. Approach :
   The approach followed to achieve the mentioned accuracy of Natural Language processing model involves the following steps:
   - **Data processing** :
     Product Category :
     - Extracted main categories from the category tree.
     - Removal of redundant data.
     - Narrow down the main categories with at least 50 examples. As the categories with fewer data would only add confusion for the model and are not enough to train the model for a particular category.

     Product Description :
   - Removing Stop word from the description
   - Stemming of description.
   - Creation of word features using the most frequently used 2000/4000 words.
   - Split the dataset into train and test : The data set is split in 75/25 ratio for training and testing purposes.

We can further use deep learning models and other techniques if there was more data available.

2. Models used: Used machine learning model:
   - SVC
   - Random Forest
   - Logistic Regression

Model Name	2000[most used words]	4000[most used words]
SVC	96.181	97.312
Random Forest	95.9924	96.60
Logistic Regression	96.4875	**97.454**

3. Result: Obtained maximum accuracy for **logistic regression** model of **97.454%.**