TEXT PREPROCESSING

Dataset:

This is an E-commerce Flipkart Dataset with exactly 20,000 samples. It has 15 columns with a lot of information. You can use is to predict which category it might fall under, considering “Description” for a product, analyse it.

Dataset Link:

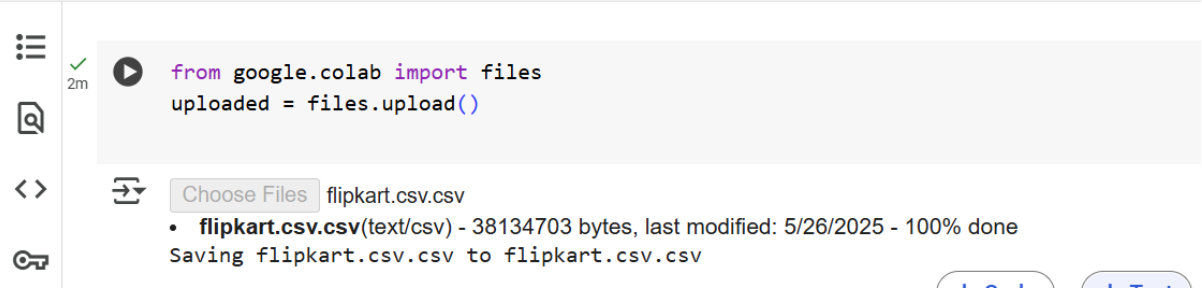
https://www.kaggle.com/datasets/atharvjairath/flipkart-ecommerce-dataset/data

Text Preprocessing Algorithm

Input: Raw product description text
Output: Cleaned and processed textual data

1. Transform all text to lowercase.
2. Eliminate all punctuation marks from the text.
3. Remove any numeric characters or digits present.
4. Split the text into individual word tokens.
5. Filter out common stop words from the token list.
6. Perform stemming to reduce words to their base/root form.
7. Apply lemmatization to convert words to their valid dictionary base forms.
8. Finally, merge the processed tokens back into a single coherent string.

Uploading CSV File:



Load the Dataset

```
import pandas as pd
df = pd.read_csv("flipkart.csv.csv")
df = df[['description']].dropna()
df.head()
```

	description
0	Key Features of Alisha Solid Women's Cycling S...
1	FabHomeDecor Fabric Double Sofa Bed (Finish Co...
2	Key Features of AW Bellies Sandals Wedges Heel...
3	Key Features of Alisha Solid Women's Cycling S...
4	Specifications of Sicons All Purpose Arnica Do...

Install and Download NLTK Resources:

```
import nltk

nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('punkt_tab')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
True
```

Define the Preprocessing Function:

```
import string
import re
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, WordNetLemmatizer

# Initialize tools
stop_words = set(stopwords.words('english'))
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()

def preprocess(text):
    # Step 1: Lowercase
    text = text.lower()

    # Step 2: Remove punctuation
    text = text.translate(str.maketrans('', '', string.punctuation))

    # Step 3: Remove numbers
    text = re.sub(r'\d+', '', text)
```

```
[5]
        # Step 4 & 5: Tokenize and remove stopwords
        tokens = word_tokenize(text)
        tokens = [word for word in tokens if word not in stop_words]

        # Step 6: Stemming
        tokens = [stemmer.stem(word) for word in tokens]

        # Step 7: Lemmatization
        tokens = [lemmatizer.lemmatize(word) for word in tokens]

        return ' '.join(tokens)
```

Apply Preprocessing:

```
[6] df['Cleaned_Description'] = df['description'].head(10).apply(preprocess)
    df[['description', 'Cleaned_Description']]
```

1 to 25 of 19998 entries Filter

index	description	Cleaned_Description
0	Key Features of Alisha Solid Women's Cycling Shorts Cotton Lycra Navy, Red, Navy,Specifications of Alisha Solid Women's Cycling Shorts Shorts Details Number of Contents in Sales Package Pack of 3 Fabric Cotton Lycra Type Cycling Shorts General Details Pattern Solid Ideal For Women's Fabric Care Gentle Machine Wash in Lukewarm Water, Do Not Bleach Additional Details Style Code ALTHT_3P_21 In the Box 3 shorts	key featur alisha solid woman cycl short cotton lycra navi red navyspecif alisha solid woman cycl short short detail number content sale packag pack fabric cotton lycra type cycl short gener detail pattern solid ideal woman fabric care gentl machin wash lukewarm water bleach addit detail style code althtp box short
	FabHomeDecor Fabric Double Sofa Bed (Finish Color - Leatherette Black Mechanism Type - Pull Out) Price: Rs. 22,646 • Fine deep seating experience • Save Space with the all new click clack Sofa Bed • Easy to fold and vice versa with simple click clack mechanism • Chrome legs with mango wood frame for long term durability • Double cushioned Sofa Bed to provide you with extra softness to make a fine seating experience • A double bed that can easily sleep two,Specifications of FabHomeDecor Fabric Double Sofa Bed (Finish Color - Leatherette Black Mechanism Type - Pull Out) Installation & Demo Installation & Demo Details Installation and demo for this product is done free of cost as part of this purchase. Our service partner will visit your location within 72 business hours from the delivery of the product. In The Box 1 Sofa Bed General Brand FabHomeDecor Mattress Included No Delivery Condition Knock Down Storage Included No Mechanism Type Pull Out Type Sofa Bed Style Contemporary & Modern	fabhomedecor fabric doubl sofa bed finish color leatherett black mechan type pull price r • fine deep seat experi • save space new click clack sofa bed • easi fold vice versa simpl click clack mechan • chrome leg mango wood frame long term durabl • doubl cushion sofa bed provid extra soft make fine seat experi • doubl bed easili sleep twospecif fabhomedecor fabric doubl sofa bed finish color leatherett black mechan type pull instal demo instal demo detail instal demo product done free cost part purchas servic partner visit locat within busi hour deliveri product box sofa bed gener brand fabhomedecor mattress includ deliveri condit knock storag includ mechan type pull type sofa bed style contemporari modern fill materi microfib seat capac seater upholsteri type na upholsteri
2	Key Features of AW Bellies Sandals Wedges Heel Casuals,AW Bellies Price: Rs. 499 Material: Synthetic Lifestyle: Casual Heel Type: Wedge Warranty Type: Manufacturer Product Warranty against manufacturing defects: 30 days Care instructions: Allow your pair of shoes to air and de-odorize at regular basis; use shoe bags to prevent any stains or mildew; dust any dry dirt from the surface using a clean cloth; do not use polish or shiner,Specifications of AW Bellies General Ideal For Women Occasion Casual Shoe Details Color Red Outer Material Patent Leather Heel Height 1 inch Number of Contents in Sales Package Pack of 1 In the Box One Pair Of Shoes	key featur aw belli sandal wedg heel casualsaw belli price r materi synthet lifestyl casual heel type wedg warranti type manufactur product warranti manufactur defect day care instruct allow pair shoe air deodor regular basi use shoe bag prevent stain mildew dust dri dirt surfac use clean cloth use polish shinerspecif aw belli gener ideal woman occas casual shoe detail color red outer materi patent leather heel height inch number content sale packag pack box one pair shoe
3	Key Features of Alisha Solid Women's Cycling Shorts Cotton Lycra Black, Red,Specifications of Alisha Solid Women's Cycling Shorts Shorts Details Number of Contents in Sales Package Pack of 2 Fabric Cotton Lycra Type Cycling Shorts General Details Pattern Solid Ideal For Women's Fabric Care Gentle Machine Wash in Lukewarm Water, Do Not Bleach Additional Details Style Code ALTGHT_11 In the Box 2 shorts	key featur alisha solid woman cycl short cotton lycra black redspecif alisha solid woman cycl short short detail number content sale packag pack fabric cotton lycra type cycl short gener detail pattern solid ideal woman fabric care gentl machin wash lukewarm water bleach addit detail style code altght box short
4	Specifications of Sicons All Purpose Arnica Dog Shampoo (500 ml) General Pet Type Dog Brand Sicons Quantity 500 ml Model Number SH.DF-14 Type All Purpose Fragrance Arnica Form Factor Liquid In the Box Sales Package Shampoo Sicons Dog Fashion Arnica	specif sicon purpos arnica dog shampoo ml gener pet type dog brand sicon quantiti ml model number shdf type purpos fragranc arnica form factor liquid box sale packag shampoo sicon dog fashion arnica
5	Key Features of Eternal Gandhi Super Series Crystal Paper Weights with Silver Finish Crystal paper weight Product Dimensions : 8cm x 8cm x 5cm A beautiful product Material: Crystal,Eternal Gandhi Super Series Crystal Paper Weights with Silver Finish (Set Of 1, Clear) Price: Rs. 430 Your office desk will sparkle and shine when you accent tables with this elegant crystal paper weight. The multifaceted crystal features Gandhiji's bust and his timeless message – "My life is my message – M.K. Gandhi". A beautiful product to gift to your near and dear ones in family and Business. Specifications of Eternal Gandhi Super Series Crystal Paper Weights with Silver Finish (Set Of 1, Clear)	key featur etern gandhi super seri crystal paper weight silver finish crystal paper weight product dimens cm x cm x cm beauti product materi crystaletern gandhi super seri crystal paper weight silver finish set clear price r offic desk sparkl shine accent tabl eleg crystal paper weight multifacet crystal featur gandhiji ' bust timeless messag – " life messag – mk gandhi " beauti product gift near dear one famili businessspecif etern gandhi super seri crystal

Variables Terminal 9:28 PM Python 3