Exclusion-Keywords-Match

October 11, 2022

0.0.1 About:

You have a list of words. You want to check if a DataFrame column contains any word from this list. Essentially , you want to exclude the rows that contain those words.

[1]:
```
import pandas as pd
import numpy as np
```

0.0.2 Read the raw data

[25]:
```
fuzz_data = pd.read_excel("TM-Data1-2021.xlsx")
fuzz_data.head(3)
```

[25]:
```
   VN_INDEX   HS_Code        Date  \
0         1  63079090  2021/01/31
1         2  63079090  2021/01/31
2         3  63071090  2021/01/31

                                    Detailed_Product
0  Dust masks made of cloth (not a medical mask),…
1  Textile yarn card strap with hooks course, bra…
2  FS-10 337 # & Tatters white cotton fabric (siz…
```

[27]:
```
fuzz_data.shape
```

[27]:
```
(309065, 4)
```

[29]:
```
raw_data = Fuzz_data.copy()
```

[30]:
```
raw_data.shape
```

[30]:
```
(309065, 4)
```

[87]:
```
#Check for duplicates
raw_data.duplicated().sum()
```

[87]:
```
0
```

[32]:
```
#converting text to lower
raw_data = raw_data.apply(lambda x: x.astype(str).str.lower())
```

Filtering out Rows in column “Detailed_Product” , containing any word from the list below:

[33]:
```
exclude_list = ['printers', 'photocopiers', 'microphone', 'memory','mobile',
'sewing', 'embroidery','telecom' , 'antenna', 'camera' , 'phone' , 'circuit' ,␣
 ↪'lawn' , 'decoration' , 'washing',
'refrigerator' , 'freezer' , 'motorcycle' , 'audio' , 'jewelry' , 'jewellery' ,␣
 ↪'vehicle' , 'furniture' , 'car',
                'decorative' , 'automotive' , 'automobile' , 'cameras' ,␣
 ↪'treadmill' , 'speakers' , 'network', 'wireless',
                'cars' , 'shredders' , 'wheels' , 'wheel' , 'Ford' , 'Steering'␣
 ↪, 'travel' , 'gearbox',
                'excavators' , 'locks'  , 'stapling', 'drilling' , 'diesel' ,␣
 ↪'truck' , 'motorbikes' , 'printer'
                'toys' , 'gaming' , 'gasoline', 'animal' , 'headphone' ,␣
 ↪'welding' , 'drills' , 'atm' , 'headset',
                'Cigarette' , 'led' , 'transformers' , 'watches' , 'toyota' ,␣
 ↪'tourist' , 'puma' , 'document',
                'scanner' , 'scanners' , 'doorbell' , 'doorbells' , 'bicycle' ,␣
 ↪'purifiers' , 'purifier',
                'screwdrivers' , 'chairs' , 'satellite', 'garden' , 'cleaner' ,␣
 ↪'women' , 'bullets', 'dishwasher',
                'scanning' , 'copier' , 'honda' , 'mounting' , 'mount' ,␣
 ↪'buttons' , 'door' , 'sanitary',
                'washers' , 'computers' , 'samsung' , 'microwave' , 'charging' ,␣
 ↪'fishing' , 'stamps' , 'labels',
                'washer' , 'manufacturing', 'propeller' , 'tweezers' ,␣
 ↪'insurance']
```

[34]:
```
#Checking the shape of the dataframe with these words
raw_data.loc[raw_data['Detailed_Product'].apply(lambda x: any([k in x for k in␣
 ↪exclude_list]))].shape
```

[34]:
```
(132099, 4)
```

[35]:
```
raw_data.head(2)
```

[35]:
```
  VN_INDEX   HS_Code        Date  \
0        1  63079090  2021/01/31
1        2  63079090  2021/01/31

                                    Detailed_Product
0  dust masks made of cloth (not a medical mask),…
1  textile yarn card strap with hooks course, bra…
```

[36]:

```
#Creating a separate df where columns do not contain the words in the list
raw_data = raw_data.loc[~raw_data['Detailed_Product'].apply(lambda x: any([k in x for k in exclude_list]))]
```

[44]:

```
raw_data.shape
```

[44]:

```
(176966, 4)
```

[45]:

```
132099  + 176966
```

[45]:

```
309065
```

We have successfully filtered out the rows from the column "Detailed_Product" where it does not contain any word from the list.

- Aisha Khalid