Untitled

January 8, 2024

[134]:
```python
import pandas as pd
import pandas as pd
import pickle
import nltk
from nltk.tokenize import RegexpTokenizer
from nltk.stem import WordNetLemmatizer,PorterStemmer
from nltk.corpus import stopwords
import re
from datasets import load_dataset
```

[135]:
```python
dataset = load_dataset('samsum')
```

```
Found cached dataset samsum (/home/sujit/.cache/huggingface/datasets/samsum/sams
um/0.0.0/f1d7c6b7353e6de335d444e424dc002ef70d1277109031327bc9cc6af5d3d46e)
100%|
                                                | 3/3
[00:00<00:00, 13.06it/s]
```

[136]:
```python
dataset
```

[136]:
```
DatasetDict({
    train: Dataset({
        features: ['id', 'dialogue', 'summary'],
        num_rows: 14732
    })
    test: Dataset({
        features: ['id', 'dialogue', 'summary'],
        num_rows: 819
    })
    validation: Dataset({
        features: ['id', 'dialogue', 'summary'],
        num_rows: 818
    })
})
```

[137]:
```python
import re
def striphtml(data):
    p = re.compile(r'<(.*)>.*?|<(.*) />')
    return p.sub('', data)

def preprocess(sentence):
    #print("setence before parsing",sentence)
    sentence=str(sentence)
    sentence = sentence.lower()
    sentence =  striphtml(sentence)
    sentence=sentence.replace('{html}',"")
    cleanr = re.compile('<.*?>')
    cleantext = re.sub(cleanr, '', sentence)
    rem_url=re.sub(r'http\S+', '',cleantext)
    rem_num = re.sub('[0-9]+', '', rem_url)
    tokenizer = RegexpTokenizer(r'\w+')
    tokens = tokenizer.tokenize(rem_num)
    filtered_words = [w for w in tokens if len(w) > 1 ]
#     filtered_words = [w for w in filtered_words if w not in stopwords.
 ↪words('english')]
    #stem_words=[stemmer.stem(w) for w in filtered_words]
#     lemma_words=[lemmatizer.lemmatize(w) for w in filtered_words]
    #print("filtered word"," ".join(lemma_words))
    return " ".join(filtered_words)
```

[ ]:

```

```

[138]:

```
dialogues = []
dia_len=[]
summaries =[]
sum_len=[]
print(len(dataset['train']))
# Extract dialogue and summary information from the dataset
for i in range(len(dataset['train'])):
    a = preprocess(dataset['train'][i]['dialogue'])
    b = preprocess(dataset['train'][i]['summary'])
    if len(a.split()) ==0 or len(b.split()) == 0 :
        continue
    else:
        dialogues.append(a)
        dia_len.append(len(a.split()))
        summaries.append(b)
        sum_len.append(len(b.split()))



print("the length of dia_len list ", len(dia_len))
print("the length of sum_len list ", len(sum_len))
```

```python

# Create a DataFrame
df = pd.DataFrame({'dialogue': dialogues, 'summary': summaries})

# Display the DataFrame
print(df.head())
df = df.reset_index(drop=True)
df.to_csv("clean_train.csv")
df.columns
```

```
14732
the length of dia_len list  14731
the length of sum_len list  14731
                                            dialogue  \
0  amanda baked cookies do you want some jerry su…
1  olivia who are you voting for in this election…
2  tim hi what up kim bad mood tbh was going to d…
3  edward rachel think in ove with bella rachel d…
4  sam hey overheard rick say something sam don k…

                                             summary
0  amanda baked cookies and will bring jerry some…
1  olivia and olivier are voting for liberals in …
2  kim may try the pomodoro technique recommended…
3  edward thinks he is in love with bella rachel …
4  sam is confused because he overheard rick comp…
```

[138]:

```
Index(['dialogue', 'summary'], dtype='object')
```

[139]:

```python
df_count = pd.DataFrame({'dial_word_count': dia_len, 'sum_word_count': sum_len})
```

[140]:

```python
x=0
for a in df_count["dial_word_count"]:
    if a== 0:
        x=x+1
print(x)
x=0
for a in df_count["sum_word_count"]:
    if a== 0:
        x=x+1
print(x)
```

```
0
0
```

[141]:

```python
x
```

```
[141]: 0
```

```
[142]: df_count.columns
```

```
[142]: Index(['dial_word_count', 'sum_word_count'], dtype='object')
```

```
[143]: df_count["sum_word_count"].describe()
```

```
[143]: count    14731.000000
        mean        19.600638
        std         10.788076
        min          1.000000
        25%         11.000000
        50%         17.000000
        75%         26.000000
        max         60.000000
        Name: sum_word_count, dtype: float64
```

```
[144]: df_count["dial_word_count"].describe()
```

```
[144]: count    14731.000000
        mean        86.693436
        std         69.118657
        min          5.000000
        25%         36.000000
        50%         67.000000
        75%        118.000000
        max        733.000000
        Name: dial_word_count, dtype: float64
```

```
[145]: dialogues = []
       dia_len=[]
       summaries =[]
       sum_len=[]
       print(len(dataset['test']))
       # Extract dialogue and summary information from the dataset
       for i in range(len(dataset['test'])):
           a = preprocess(dataset['test'][i]['dialogue'])
           b = preprocess(dataset['test'][i]['summary'])
           if len(a.split()) ==0 or len(b.split()) == 0 :
               continue
           else:
               dialogues.append(a)
               dia_len.append(len(a.split()))
               summaries.append(b)
               sum_len.append(len(b.split()))
```

```python


print("the length of dia_len list ", len(dia_len))
print("the length of sum_len list ", len(sum_len))




# Create a DataFrame
df = pd.DataFrame({'dialogue': dialogues, 'summary': summaries})

# Display the DataFrame
print(df.head())
df = df.reset_index(drop=True)
df.to_csv("clean_test.csv")
df.columns
```

```
819
the length of dia_len list  819
the length of sum_len list  819
                                            dialogue  \
0  hannah hey do you have betty number amanda lem…
1  eric machine rob that so gr eric know and show…
2  lenny babe can you help me with something bob …
3  will hey babe what do you want for dinner toni…
4  ollie hi are you in warsaw jane yes just back …

                                             summary
0  hannah needs betty number but amanda doesn hav…
1  eric and rob are going to watch stand up on yo…
2  lenny can decide which trousers to buy bob adv…
3  emma will be home soon and she will let will know
4  jane is in warsaw ollie and jane has party jan…
```

[145]:

```
Index(['dialogue', 'summary'], dtype='object')
```

[146]:

```python
df_count = pd.DataFrame({'dial_word_count': dia_len, 'sum_word_count': sum_len})
```

[147]:

```python
x=0
for a in df_count["dial_word_count"]:
    if a== 0:
        x=x+1
print(x)
x=0
for a in df_count["sum_word_count"]:
    if a== 0:
        x=x+1
print(x)
```

```
0
0
```

[148]:

```
df_count["dial_word_count"].describe()
```

[148]:

```
count    819.000000
mean      88.096459
std       69.969129
min        6.000000
25%       38.000000
50%       68.000000
75%      117.000000
max      501.000000
Name: dial_word_count, dtype: float64
```

[149]:

```
df_count["sum_word_count"].describe()
```

[149]:

```
count    819.000000
mean      19.268620
std       10.300076
min        3.000000
25%       11.000000
50%       17.000000
75%       25.000000
max       56.000000
Name: sum_word_count, dtype: float64
```

[150]:

```
dialogues = []
dia_len=[]
summaries =[]
sum_len=[]
print(len(dataset['validation']))
# Extract dialogue and summary information from the dataset
for i in range(len(dataset['validation'])):
    a = preprocess(dataset['validation'][i]['dialogue'])
    b = preprocess(dataset['validation'][i]['summary'])
    if len(a.split()) ==0 or len(b.split()) == 0 :
        continue
    else:
        dialogues.append(a)
        dia_len.append(len(a.split()))
        summaries.append(b)
        sum_len.append(len(b.split()))



print("the length of dia_len list ", len(dia_len))
```

```python
print("the length of sum_len list ", len(sum_len))



# Create a DataFrame
df = pd.DataFrame({'dialogue': dialogues, 'summary': summaries})

# Display the DataFrame
print(df.head())
df = df.reset_index(drop=True)
df.to_csv("clean_valid.csv")
df.columns
```

```
818
the length of dia_len list  818
the length of sum_len list  818
                                            dialogue  \
0  hi tom are you busy tomorrow afternoon pretty …
1  emma ve just fallen in love with this advent c…
2  jackie madison is pregnant jackie but she does…
3  marla marla look what found under my bed kiki …
4  robert hey give me the address of this music s…

                                             summary
0  will go to the animal shelter tomorrow to get …
1  emma and rob love the advent calendar lauren f…
2  madison is pregnant but she doesn want to talk…
3           marla found pair of boxers under her bed
4  robert wants fred to send him the address of t…
```

[150]:

```
Index(['dialogue', 'summary'], dtype='object')
```

[151]:

```python
df_count = pd.DataFrame({'dial_word_count': dia_len, 'sum_word_count': sum_len})
```

[152]:

```python
x=0
for a in df_count["dial_word_count"]:
    if a== 0:
        x=x+1
print(x)
x=0
for a in df_count["sum_word_count"]:
    if a== 0:
        x=x+1
print(x)
```

```
0
0
```

[153]:
```
df_count["dial_word_count"].describe()
```

[153]:
```
count    818.000000
mean      84.672372
std       69.511305
min        9.000000
25%       35.000000
50%       65.000000
75%      117.000000
max      508.000000
Name: dial_word_count, dtype: float64
```

[154]:
```
df_count["sum_word_count"].describe()
```

[154]:
```
count    818.000000
mean      19.537897
std       10.796058
min        3.000000
25%       11.250000
50%       17.000000
75%       26.000000
max       56.000000
Name: sum_word_count, dtype: float64
```

[ ]:

[ ]:

[ ]:

[ ]: