NER-SA-Dataframe

March 11, 2023

1 Creating Dataframe for targetted Sentiment Analysis

[10]:
```
import pandas as pd
news_df=pd.read_parquet("news_v3.parquet")
```

[11]:
```
news_df.head(2)
```

[11]:
```
   index                                                url       date  \
0      0  http://auckland.scoop.co.nz/2020/01/aut-boosts… 2020-01-28
1      1  http://en.people.cn/n3/2021/0318/c90000-983012… 2021-03-18

  language                                              title  \
0       en  auckland.scoop.co.nz » AUT boosts AI expertise…
1       en  Artificial intelligence improves parking effic…

                                                 text  \
0  \n\nauckland.scoop.co.nz » AUT boosts AI exper…
1  \n\nArtificial intelligence improves parking e…

                                      text_cleaned  topic Sentiment
0  aucklandscoopconz aut boost ai expertise new a…      3  Positive
1  artificial intelligence improves parking effic…      4  Negative
```

[12]:
```
news_df.shape
```

[12]:
```
(199838, 9)
```

[13]:
```
org_list=["youtube", "chatgpt", "samsung" , "microsoft" , "google" , "gray␣
 ↪television" , "gray medium" ,  "nexstar" , "ibm", "nasa" , "fcc" ,"intel" ]

loc_list=["new york", "india", "china" ,"japan","california" ,"washington" ,␣
 ↪"australia" , "canada" , "israel" , "germany" , "france" , "london" , "uk" ,␣
 ↪"singapore"  , "pennsylvania"  , "doylestown" , "allenstown"]
```

[14]:
```
# Define column names
columns = ['date', 'topic', 'sentiment' , 'ent','ent_val', 'sent']
```

```
# Create empty dataframe with columns
ner_sa = pd.DataFrame(columns=columns)
```

[18]:
```
import pandas as pd
import nltk
from nltk import word_tokenize, pos_tag
import spacy
from nltk.tokenize import word_tokenizec
```

```
2023-03-10 22:49:39.132823: I tensorflow/core/platform/cpu_feature_guard.cc:193]
This TensorFlow binary is optimized with oneAPI Deep Neural Network Library
(oneDNN) to use the following CPU instructions in performance-critical
operations:  AVX2 FMA
To enable them in other operations, rebuild TensorFlow with the appropriate
compiler flags.
2023-03-10 22:49:43.281454: W
tensorflow/compiler/xla/stream_executor/platform/default/dso_loader.cc:64] Could
not load dynamic library 'libcudart.so.11.0'; dlerror: libcudart.so.11.0: cannot
open shared object file: No such file or directory
2023-03-10 22:49:43.281484: I
tensorflow/compiler/xla/stream_executor/cuda/cudart_stub.cc:29] Ignore above
cudart dlerror if you do not have a GPU set up on your machine.
2023-03-10 22:49:52.113600: W
tensorflow/compiler/xla/stream_executor/platform/default/dso_loader.cc:64] Could
not load dynamic library 'libnvinfer.so.7'; dlerror: libnvinfer.so.7: cannot
open shared object file: No such file or directory
2023-03-10 22:49:52.115066: W
tensorflow/compiler/xla/stream_executor/platform/default/dso_loader.cc:64] Could
not load dynamic library 'libnvinfer_plugin.so.7'; dlerror:
libnvinfer_plugin.so.7: cannot open shared object file: No such file or
directory
2023-03-10 22:49:52.115086: W
tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Cannot
dlopen some TensorRT libraries. If you would like to use Nvidia GPU with
TensorRT, please make sure the missing libraries mentioned above are installed
properly.
2023-03-10 22:49:57.447314: W
tensorflow/compiler/xla/stream_executor/platform/default/dso_loader.cc:64] Could
not load dynamic library 'libcuda.so.1'; dlerror: libcuda.so.1: cannot open
shared object file: No such file or directory
2023-03-10 22:49:57.464554: W
tensorflow/compiler/xla/stream_executor/cuda/cuda_driver.cc:265] failed call to
cuInit: UNKNOWN ERROR (303)
2023-03-10 22:49:57.464633: I
tensorflow/compiler/xla/stream_executor/cuda/cuda_diagnostics.cc:156] kernel
driver does not appear to be running on this host (python-20230307-192621):
/proc/driver/nvidia/version does not exist
```

[19]:
```
nlp = spacy.load("en_core_web_lg")
```

[24]:
```
for i in range(len(news_df)):
    if i % 2000 == 0:
        print(i)
    for sent in nltk.sent_tokenize(news_df['text_cleaned'][i]):
        #doc = list(nlp(sent))
        for org in org_list:
            if org in sent:
                ␣
 ↪new_row=[news_df['date'][i],news_df['topic'][i],news_df['Sentiment'][i],'ORG'␣
 ↪, org , doc]
                ner_sa = ner_sa.append(pd.Series(new_row, index=columns),␣
 ↪ignore_index=True)
        for loc in loc_list:
            if loc in sent:
                ␣
 ↪new_row=[news_df['date'][i],news_df['topic'][i],news_df['Sentiment'][i],'LOC'␣
 ↪, loc , doc]
                ner_sa = ner_sa.append(pd.Series(new_row, index=columns),␣
 ↪ignore_index=True)
```

```
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000
32000
34000
36000
38000
40000
42000
44000
46000
48000
```

50000
52000
54000
56000
58000
60000
62000
64000
66000
68000
70000
72000
74000
76000
78000
80000
82000
84000
86000
88000
90000
92000
94000
96000
98000
100000
102000
104000
106000
108000
110000
112000
114000
116000
118000
120000
122000
124000
126000
128000
130000
132000
134000
136000
138000
140000
142000
144000

```
146000
148000
150000
152000
154000
156000
158000
160000
162000
164000
166000
168000
170000
172000
174000
176000
178000
180000
182000
184000
186000
188000
190000
192000
194000
196000
198000
```

[25]:
```
ner_sa.head(2)
```

[25]:
```
        date topic sentiment  ent ent_val  \
0 2020-01-28     3  Positive  ORG   intel
1 2020-01-28     3  Positive  ORG   intel

                                                sent
0  [aucklandscoopconz, aut, boost, ai, expertise,…
1  [aucklandscoopconz, aut, boost, ai, expertise,…
```

[26]:
```
ner_sa.shape
```

[26]:
```
(994210, 6)
```

[28]:
```
ner_sa.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 994210 entries, 0 to 994209
Data columns (total 6 columns):
 #   Column     Non-Null Count   Dtype
```

```
---  ------     --------------   -----
 0   date       994210 non-null  datetime64[ns]
 1   topic      994210 non-null  object
 2   sentiment  994210 non-null  object
 3   ent        994210 non-null  object
 4   ent_val    994210 non-null  object
 5   sent       994210 non-null  object
dtypes: datetime64[ns](1), object(5)
memory usage: 45.5+ MB
```

[29]:
```
ner_sa.to_csv("ner_sa_df.csv")
```

[31]:
```
df=pd.read_csv("ner_sa_df.csv")
df.shape
```

[31]:
```
(994210, 7)
```

[32]:
```
df.head(2)
```

[32]:
```
   Unnamed: 0        date  topic sentiment  ent ent_val  \
0           0  2020-01-28      3  Positive  ORG   intel
1           1  2020-01-28      3  Positive  ORG   intel

                                                sent
0  [aucklandscoopconz, aut, boost, ai, expertise,…
1  [aucklandscoopconz, aut, boost, ai, expertise,…
```

[ ]: