Untitled

June 11, 2022

[1]:
```
import utility as util
import pandas as pd
import os
import numpy as np
import gzip
import time
```

[2]:
```
#Input file: en-fr.csv    Size:8.2GB    Source:https://www.kaggle.com/datasets/
 ↪dhruvildave/en-fr-translation-dataset?resource=download
```

[4]:
```
%%writefile config.yaml
file_type: csv
dataset_name: testfile
file_name: en-fr
inbound_delim: ","
outbound_delim: "|"
columns:
    - fr
    - en
```

```
Overwriting config.yaml
```

[5]:
```
#load config file
cfg = util.read_cfg("config.yaml")
```

[6]:
```
cfg
```

[6]:
```
{'file_type': 'csv',
 'dataset_name': 'testfile',
 'file_name': 'en-fr',
 'inbound_delim': ',',
 'outbound_delim': '|',
 'columns': ['fr', 'en']}
```

[7]:
```
#traditional way to read file
start = time.perf_counter()
df_pd = pd.read_csv("en-fr.csv")
end = time.perf_counter()
```

```python
print("File loading time using Panda in seconds: " + str(end - start))
del df_pd
```

```
File loading time using Panda in seconds: 88.716633
```

[8]:

```python
#Try to read file using Dask
import dask.dataframe as dd
start = time.perf_counter()
df_dask = dd.read_csv("en-fr.csv")
end = time.perf_counter()
print("File loading time using Dask in seconds: " + str(end - start))
del df_dask
```

```
File loading time using Dask in seconds: 0.03396560000000193
```

[9]:

```python
#Try to read file using Ray
import ray
start = time.perf_counter()
df_ray = ray.data.read_csv("en-fr.csv")
end = time.perf_counter()
print("File loading time using Ray in seconds: " + str(end - start))
del df_ray
```

```
2022-06-11 07:12:30,528 WARNING read_api.py:252 -- The number of blocks in this
dataset (1) limits its parallelism to 1 concurrent tasks. This is much less than
the number of available CPU slots in the cluster. Use `.repartition(n)` to
increase the number of dataset blocks.
```

```
File loading time using Ray in seconds: 21.49454250000001
```

[10]:

```python
#Use config file to read
source_file = "./" + cfg["file_name"] + "."+ cfg["file_type"]
df = pd.read_csv(source_file, cfg["inbound_delim"])
df.head()
```

```
C:\Users\songz\anaconda3\lib\site-
packages\IPython\core\interactiveshell.py:3444: FutureWarning: In a future
version of pandas all arguments of read_csv except for the argument
'filepath_or_buffer' will be keyword-only
  exec(code_obj, self.user_global_ns, self.user_ns)
```

[10]:

```
                                                 en  \
0  Changing Lives | Changing Society | How It Wor…
1                                          Site map
2                                          Feedback
3                                           Credits
4                                          Français
```

```
                                                   fr
0  Il a transformé notre vie | Il a transformé la…
1                                        Plan du site
2                                         Rétroaction
3                                             Crédits
4                                             English
```

[11]:

```
#Calculate file statistic
row_count = len(df)
col_count = len(df.columns)
```

[12]:

```
#Validation and output gz file
if util.col_header_val(df, cfg):
    print("validation pass")
    outfile_name = cfg["dataset_name"] + ".txt.gz"
    df.to_csv('temp.txt', index=False, sep= cfg["outbound_delim"])
    with open("temp.txt", 'rb') as orig_file:
        with gzip.open(outfile_name, 'wb') as zipped_file:
            zipped_file.writelines(orig_file)
    os.remove('temp.txt')
    file_size = os.path.getsize("./"+ outfile_name)
    print("Total number of rows: " + str(row_count) + "    Total number of␣
 ↪columns: " + str(col_count) + "    Output file size: " + str(file_size) + "␣
 ↪Byte")
else:
    print("validation fail")
```

```
column name validation passed
validation pass
Total number of rows: 22520376    Total number of columns: 2    Output file
size: 2668521476 Byte
```

[ ]:

```

```