Тестовое задание на анализ данных.

Елизавета Рыжова

```
import pandas as pd
import matplotlib.pyplot as plt
import os
```

```
!chcp 1251
```

```
Текущая кодовая страница: 1251
```

Предположим, что архив находится в рабочей директории.

```
if not "exmpl" in os.listdir():
    !wsl unzip -q exmpl.zip -d ./exmpl
```

Поскольку данные имеют ошибки с длинной строк и форматом, будем загружать данные по строчкам.
Сложим все данные в один Датафрейм.
Сделаем индексацию по 'datetime' извлечённой из полей с датами и временем.

```
def read_file(path):
    with open(path, 'r', encoding='latin1') as file:
        lines = file.readlines()

    counter = 0
    rows = []
    for line in lines[3:]:  # skip the header and empty rows
        row = line.strip().split(' ')
        if len(row) == 15:
            rows.append(row)
        else:
            counter += 1
    name = path.split('/')[-1]
    print(f'Файл {name} \nПотеряно строк {counter}')
    df = pd.DataFrame(rows, columns=['Date', 'Time', 's-ip', 'cs-method', 'cs

    df['datetime'] = pd.to_datetime(df['Date'] + ' ' + df['Time'])
    return df
```

```
file_names = os.listdir('./exmpl')

dfs = []
```

```
 4
 5 for file_name in file_names:
 6     df = read_file('./exmpl/' + file_name)
 7     df.info()
 8     display(df.head(3))
 9     dfs.append(df)
10
11 logs_df = pd.concat(dfs)
12 logs_df.set_index('datetime', inplace=True)
```

```
Файл u_ex200328.log
Потеряно строк 5
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2608910 entries, 0 to 2608909
Data columns (total 16 columns):
 #   Column           Dtype
---  ------           -----
 0   Date             object
 1   Time             object
 2   s-ip             object
 3   cs-method        object
 4   cs-uri-stem      object
 5   cs-uri-query     object
 6   s-port           object
 7   cs-username      object
 8   c-ip             object
 9   cs(User-Agent)   object
 10  cs(Referer)      object
 11  sc-status        object
 12  sc-substatus     object
 13  sc-win32-status  object
 14  time-taken       object
 15  datetime         datetime64[ns]
dtypes: datetime64[ns](1), object(15)
memory usage: 318.5+ MB
```

	Date	Time	s-ip	cs-method	
0	2020-03-28	00:00:00	192.168.254.71	POST	/1c_askona_
1	2020-03-28	00:00:00	192.168.254.71	POST	/1c_askona_
2	2020-03-28	00:00:00	192.168.254.71	POST	/1c_askona_

```
Файл u_ex200723.log
Потеряно строк 9
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129684 entries, 0 to 129683
Data columns (total 16 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   Date             129684 non-null  object
 1   Time             129684 non-null  object
 2   s-ip             129684 non-null  object
 3   cs-method        129684 non-null  object
 4   cs-uri-stem      129684 non-null  object
 5   cs-uri-query     129684 non-null  object
 6   s-port           129684 non-null  object
 7   cs-username      129684 non-null  object
 8   c-ip             129684 non-null  object
 9   cs(User-Agent)   129684 non-null  object
 10  cs(Referer)      129684 non-null  object
 11  sc-status        129684 non-null  object
```

Данные загружены и осмотрены.
Изменим названия колонок и тип данных.

```
1 logs_df.rename(columns=lambda x: x.replace('-', '_'), inplace=True)
```

```
1 logs_df[['s_port', 'sc_status', 'time_taken']] = logs_df[['s_port', 'sc_statu
```

```
1 logs_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 3702214 entries, 2020-03-28 00:00:00 to 2020-07-30 23:59:50
Data columns (total 15 columns):
 #   Column           Dtype 
---  ------           ----- 
 0   Date             object
 1   Time             object
 2   s_ip             object
 3   cs_method        object
 4   cs_uri_stem      object
 5   cs_uri_query     object
 6   s_port           int32 
 7   cs_username      object
 8   c_ip             object
 9   cs(User_Agent)   object
 10  cs(Referer)      object
 11  sc_status        int32 
 12  sc_substatus     object
 13  sc_win32_status  object
 14  time_taken       int32 
dtypes: int32(3), object(12)
memory usage: 409.6+ MB
```

Представлены данные с 28 Марта 20го года по 30 Июля 20го года.

```
1 logs_df.s_ip.unique()
```

```
array(['192.168.254.71'], dtype=object)
```

s_ip во всех наблюдениях один и тот же.

```
1 logs_df.cs_uri_stem.unique()
```

```
array(['/1c_askona_wms_2/ws/galESBExchange.1cws',
       '/1c_askona_wms_msk-1/ws/galESBExchange.1cws',
       '/erp_pmt_repl_signer1/hs/dadir/getEmployees', ...,
       '/1c-ask-wms-obuhovo-test/ru_RU/e1csys/mngsrv/empty.gif',
       '/e1csys/mngsrv/favicon.ico', '/telephony-service.html'],
      dtype=object)
```

```
1 logs_df.c_ip.unique()
```