S04_T01

May 12, 2022

0.0.1 Nivell 1

L'analista ha d'assegurar-se que els registres consisteixen en una gamma completa de missatges i s'interpreten segons el context. Els elements de registre han d'estandaritzar-se, utilitzant els mateixos termes o terminologia, per evitar confusions i proporcionar cohesió.

Com Científic de Dades se t'ha proporcionat accés als registres-Logs on queda registrada l'activitat de totes les visites a realitzades a la pàgina web de l'agència de viatges "akumenius.com".

- Exercici 1

Estandaritza, identifica i enumera cada un dels atributs / variables de l'estructura de l'arxiu "Web_access_log-akumenius.com" que trobaràs al repositori de GitHub "Data-sources".

- Exercici 2

Neteja, preprocesa, estructura i transforma (dataframe) les dades del registre d'Accés a la web.

```
[1]: import csv
import re
import pandas as pd
import numpy as np
import seaborn as sb
import matplotlib.pyplot as plt
import matplotlib.ticker as mtick
from scipy import stats


sep = r'^(\S+)\s(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})\s-\s-\s\[(.
 ↪+)\]\s\"(\S+)\s(\S+)\s(\S+)\"\s(\S+).+\s\"(\S+)\s\((.+)\).+$'
header =␣
 ↪['web','client_IP','timestamp','event','consulta','protocol','http_status_code','explorer',
 ↪'user_agent','rest']

df = pd.read_csv(r'C:
 ↪\Users\motxi\Documents\Data_Science_IT_Academy\Sprints\S04\Web_access_log-akumenius.
 ↪com.txt',
                 engine = 'python',
                 delimiter = sep,
                 names = header,
                 encoding = 'ISO-8859-1')
```

```
pd.to_datetime(df['timestamp'], format='%dd/%MMM/%yyyy:%HH:%mm:%SS', utc=True,␣
↪errors='ignore' )

log_df = df.
↪filter(['web','client_IP','timestamp','event','consulta','protocol','http_status_code','exp
↪axis=1)

log_df.head()
```

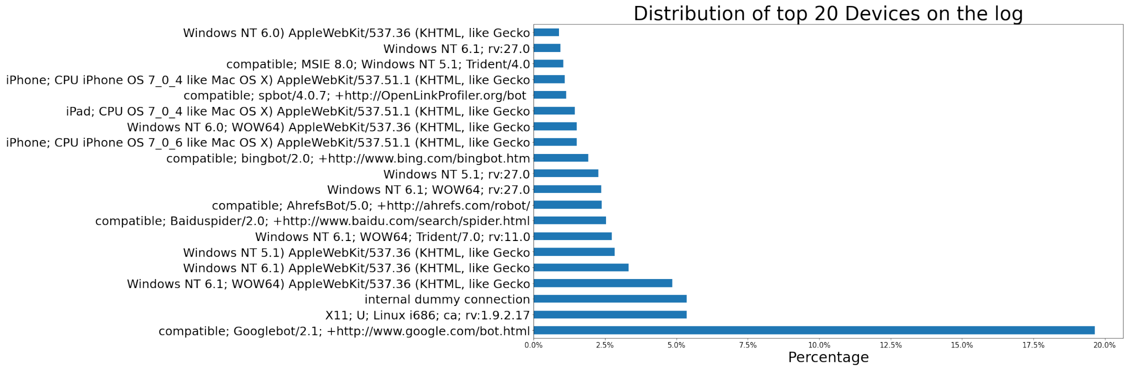
```
[1]:             web  client_IP                   timestamp    event consulta  \
NaN  localhost  127.0.0.1  23/Feb/2014:03:10:31 +0100  OPTIONS        *
NaN  localhost  127.0.0.1  23/Feb/2014:03:10:31 +0100  OPTIONS        *
NaN  localhost  127.0.0.1  23/Feb/2014:03:10:31 +0100  OPTIONS        *
NaN  localhost  127.0.0.1  23/Feb/2014:03:10:31 +0100  OPTIONS        *
NaN  localhost  127.0.0.1  23/Feb/2014:03:10:31 +0100  OPTIONS        *

     protocol  http_status_code explorer                 user_agent  rest
NaN  HTTP/1.0             200.0   Apache  internal dummy connection   NaN
NaN  HTTP/1.0             200.0   Apache  internal dummy connection   NaN
NaN  HTTP/1.0             200.0   Apache  internal dummy connection   NaN
NaN  HTTP/1.0             200.0   Apache  internal dummy connection   NaN
NaN  HTTP/1.0             200.0   Apache  internal dummy connection   NaN
```

```
[2]: print(log_df.mode(dropna=True), '\n')
```

```
                web      client_IP                   timestamp event  \
0  www.akumenius.com  66.249.76.216  28/Feb/2014:04:16:25 +0100   GET

  consulta  protocol  http_status_code     explorer  \
0        *  HTTP/1.1             200.0  Mozilla/5.0

                                        user_agent  rest
0  compatible; Googlebot/2.1; +http://www.google…   NaN
```

```
[3]: fig,ax = plt.subplots(figsize = (22,12))
     print(log_df['user_agent'].value_counts(normalize=True).mul(100)[:10])

     ax = log_df['user_agent'].value_counts(normalize=True).mul(100)[:20].
      ↪plot(kind='barh')
     ax.tick_params(axis='y', labelsize=25)
     ax.tick_params(axis='x', labelsize=15)
     ax.set_xlabel("Percentage", size = 30)
     ax.set_title("Distribution of top 20 Devices on the log", size = 40)
     ax.xaxis.set_major_formatter(mtick.PercentFormatter())
     plt.savefig('./Devices_travel_ekumenus.png')
     plt.show(fig)

     log_df['event'].unique()
```

```
compatible; Googlebot/2.1; +http://www.google.com/bot.html
19.658295
X11; U; Linux i686; ca; rv:1.9.2.17
5.364196
internal dummy connection
5.363810
Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko
4.858395
Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko
3.326319
Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko
2.838665
Windows NT 6.1; WOW64; Trident/7.0; rv:11.0
2.734416
compatible; Baiduspider/2.0; +http://www.baidu.com/search/spider.html
2.525917
compatible; AhrefsBot/5.0; +http://ahrefs.com/robot/
2.379969
Windows NT 6.1; WOW64; rv:27.0
2.373019
Name: user_agent, dtype: float64
```



[3]: array(['OPTIONS', 'GET', None, 'POST', 'HEAD'], dtype=object)

[4]:
```
print(log_df['timestamp'].value_counts()[:200])
```

```
28/Feb/2014:04:16:25 +0100    83
25/Feb/2014:18:01:20 +0100    76
25/Feb/2014:18:36:30 +0100    68
25/Feb/2014:15:58:34 +0100    67
```

```
26/Feb/2014:17:36:25 +0100    67
                              ..
24/Feb/2014:14:22:49 +0100    37
27/Feb/2014:16:22:06 +0100    37
25/Feb/2014:17:52:35 +0100    37
23/Feb/2014:20:07:17 +0100    37
27/Feb/2014:13:06:39 +0100    37
Name: timestamp, Length: 200, dtype: int64
```

- Exercici 3

Geolocalitza les IP's.

[5]:
```python
import socket
from requests import get
import sys
!{sys.executable} -m pip install ip2geotools
from ip2geotools.databases.noncommercial import DbIpCity as locDb #
```

```
Requirement already satisfied: ip2geotools in c:\users\motxi\anaconda3\lib\site-
packages (0.1.6)
Requirement already satisfied: pip-review>=1.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.1.1)
Requirement already satisfied: requests>=2.20.1 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (2.26.0)
Requirement already satisfied: dicttoxml>=1.7.4 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.7.4)
Requirement already satisfied: typed-ast>=1.1.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.4.3)
Requirement already satisfied: six>=1.11.0 in c:\users\motxi\anaconda3\lib\site-
packages (from ip2geotools) (1.16.0)
Requirement already satisfied: requests-toolbelt>=0.8.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (0.9.1)
Requirement already satisfied: mccabe>=0.6.1 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (0.6.1)
Requirement already satisfied: tqdm>=4.28.1 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (4.62.3)
Requirement already satisfied: astroid>=2.1.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (2.6.6)
Requirement already satisfied: selenium>=3.141.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (4.1.3)
Requirement already satisfied: isort>=4.3.4 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (5.9.3)
Requirement already satisfied: maxminddb>=1.4.1 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (2.2.0)
Requirement already satisfied: twine>=1.12.1 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (4.0.0)
Requirement already satisfied: Click>=7.0 in c:\users\motxi\anaconda3\lib\site-
packages (from ip2geotools) (8.0.3)
```

```
Requirement already satisfied: readme-renderer>=24.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (35.0)
Requirement already satisfied: ratelim>=0.1.6 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (0.1.6)
Requirement already satisfied: wrapt>=1.10.11 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.12.1)
Requirement already satisfied: chardet>=3.0.4 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (4.0.0)
Requirement already satisfied: pyquery>=1.4.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.4.3)
Requirement already satisfied: bleach>=3.0.2 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (4.0.0)
Requirement already satisfied: typing>=3.6.6 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (3.7.4.3)
Requirement already satisfied: future>=0.17.1 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (0.18.2)
Requirement already satisfied: pyparsing>=2.3.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (3.0.4)
Requirement already satisfied: autopep8>=1.4.3 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.5.7)
Requirement already satisfied: pycodestyle>=2.4.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (2.7.0)
Requirement already satisfied: decorator>=4.3.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (5.1.0)
Requirement already satisfied: webencodings>=0.5.1 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (0.5.1)
Requirement already satisfied: geocoder>=1.38.1 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.38.1)
Requirement already satisfied: Pygments>=2.3.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (2.10.0)
Requirement already satisfied: IP2Location>=8.0.3 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (8.7.3)
Requirement already satisfied: certifi>=2018.10.15 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (2021.10.8)
Requirement already satisfied: urllib3>=1.24.1 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.26.7)
Requirement already satisfied: lxml>=4.2.5 in c:\users\motxi\anaconda3\lib\site-
packages (from ip2geotools) (4.6.3)
Requirement already satisfied: pkginfo>=1.4.2 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.8.2)
Requirement already satisfied: idna>=2.7 in c:\users\motxi\anaconda3\lib\site-
packages (from ip2geotools) (3.2)
Requirement already satisfied: docutils>=0.14 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (0.17.1)
Requirement already satisfied: lazy-object-proxy>=1.3.1 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.6.0)
Requirement already satisfied: cssselect>=1.0.3 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (1.1.0)
```

```
Requirement already satisfied: packaging>=18.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (21.0)
Requirement already satisfied: geoip2>=2.9.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (4.5.0)
Requirement already satisfied: pylint>=2.2.0 in
c:\users\motxi\anaconda3\lib\site-packages (from ip2geotools) (2.9.6)
Requirement already satisfied: setuptools>=20.0 in
c:\users\motxi\anaconda3\lib\site-packages (from astroid>=2.1.0->ip2geotools)
(58.0.4)
Requirement already satisfied: toml in c:\users\motxi\anaconda3\lib\site-
packages (from autopep8>=1.4.3->ip2geotools) (0.10.2)
Requirement already satisfied: colorama in c:\users\motxi\anaconda3\lib\site-
packages (from Click>=7.0->ip2geotools) (0.4.4)
Requirement already satisfied: aiohttp<4.0.0,>=3.6.2 in
c:\users\motxi\anaconda3\lib\site-packages (from geoip2>=2.9.0->ip2geotools)
(3.8.1)
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in
c:\users\motxi\anaconda3\lib\site-packages (from
aiohttp<4.0.0,>=3.6.2->geoip2>=2.9.0->ip2geotools) (4.0.2)
Requirement already satisfied: yarl<2.0,>=1.0 in
c:\users\motxi\anaconda3\lib\site-packages (from
aiohttp<4.0.0,>=3.6.2->geoip2>=2.9.0->ip2geotools) (1.7.2)
Requirement already satisfied: aiosignal>=1.1.2 in
c:\users\motxi\anaconda3\lib\site-packages (from
aiohttp<4.0.0,>=3.6.2->geoip2>=2.9.0->ip2geotools) (1.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in
c:\users\motxi\anaconda3\lib\site-packages (from
aiohttp<4.0.0,>=3.6.2->geoip2>=2.9.0->ip2geotools) (1.3.0)
Requirement already satisfied: attrs>=17.3.0 in
c:\users\motxi\anaconda3\lib\site-packages (from
aiohttp<4.0.0,>=3.6.2->geoip2>=2.9.0->ip2geotools) (21.2.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
c:\users\motxi\anaconda3\lib\site-packages (from
aiohttp<4.0.0,>=3.6.2->geoip2>=2.9.0->ip2geotools) (6.0.2)
Requirement already satisfied: charset-normalizer<3.0,>=2.0 in
c:\users\motxi\anaconda3\lib\site-packages (from
aiohttp<4.0.0,>=3.6.2->geoip2>=2.9.0->ip2geotools) (2.0.4)
Requirement already satisfied: pip in c:\users\motxi\anaconda3\lib\site-packages
(from pip-review>=1.0->ip2geotools) (21.2.4)
Requirement already satisfied: trio-websocket~=0.9 in
c:\users\motxi\anaconda3\lib\site-packages (from selenium>=3.141.0->ip2geotools)
(0.9.2)
Requirement already satisfied: trio~=0.17 in c:\users\motxi\anaconda3\lib\site-
packages (from selenium>=3.141.0->ip2geotools) (0.20.0)
Requirement already satisfied: cffi>=1.14 in c:\users\motxi\anaconda3\lib\site-
packages (from trio~=0.17->selenium>=3.141.0->ip2geotools) (1.14.6)
Requirement already satisfied: async-generator>=1.9 in
c:\users\motxi\anaconda3\lib\site-packages (from
trio~=0.17->selenium>=3.141.0->ip2geotools) (1.10)
Requirement already satisfied: outcome in c:\users\motxi\anaconda3\lib\site-
packages (from trio~=0.17->selenium>=3.141.0->ip2geotools) (1.1.0)
Requirement already satisfied: sniffio in c:\users\motxi\anaconda3\lib\site-
packages (from trio~=0.17->selenium>=3.141.0->ip2geotools) (1.2.0)
Requirement already satisfied: sortedcontainers in
c:\users\motxi\anaconda3\lib\site-packages (from
trio~=0.17->selenium>=3.141.0->ip2geotools) (2.4.0)
Requirement already satisfied: pycparser in c:\users\motxi\anaconda3\lib\site-
packages (from cffi>=1.14->trio~=0.17->selenium>=3.141.0->ip2geotools) (2.20)
Requirement already satisfied: wsproto>=0.14 in
c:\users\motxi\anaconda3\lib\site-packages (from trio-
websocket~=0.9->selenium>=3.141.0->ip2geotools) (1.1.0)
Requirement already satisfied: importlib-metadata>=3.6 in
c:\users\motxi\anaconda3\lib\site-packages (from twine>=1.12.1->ip2geotools)
(4.8.1)
Requirement already satisfied: rfc3986>=1.4.0 in
c:\users\motxi\anaconda3\lib\site-packages (from twine>=1.12.1->ip2geotools)
(2.0.0)
Requirement already satisfied: keyring>=15.1 in
c:\users\motxi\anaconda3\lib\site-packages (from twine>=1.12.1->ip2geotools)
(23.1.0)
Requirement already satisfied: rich>=12.0.0 in
c:\users\motxi\anaconda3\lib\site-packages (from twine>=1.12.1->ip2geotools)
(12.2.0)
Requirement already satisfied: zipp>=0.5 in c:\users\motxi\anaconda3\lib\site-
packages (from importlib-metadata>=3.6->twine>=1.12.1->ip2geotools) (3.6.0)
Requirement already satisfied: pywin32-ctypes!=0.1.0,!=0.1.1 in
c:\users\motxi\anaconda3\lib\site-packages (from
keyring>=15.1->twine>=1.12.1->ip2geotools) (0.2.0)
Requirement already satisfied: commonmark<0.10.0,>=0.9.0 in
c:\users\motxi\anaconda3\lib\site-packages (from
rich>=12.0.0->twine>=1.12.1->ip2geotools) (0.9.1)
Requirement already satisfied: pyOpenSSL>=0.14 in
c:\users\motxi\anaconda3\lib\site-packages (from urllib3>=1.24.1->ip2geotools)
(21.0.0)
Requirement already satisfied: cryptography>=1.3.4 in
c:\users\motxi\anaconda3\lib\site-packages (from urllib3>=1.24.1->ip2geotools)
(3.4.8)
Requirement already satisfied: PySocks!=1.5.7,<2.0,>=1.5.6 in
c:\users\motxi\anaconda3\lib\site-packages (from urllib3>=1.24.1->ip2geotools)
(1.7.1)
Requirement already satisfied: h11<1,>=0.9.0 in
c:\users\motxi\anaconda3\lib\site-packages (from wsproto>=0.14->trio-
websocket~=0.9->selenium>=3.141.0->ip2geotools) (0.13.0)
```

[12]:

```
def createModel(db, ip): # create a response model holding attributes and␣
 ↪methods of location
    model = db.get(ip, api_key='Free')
    return model
```

[13]:

```
def extractModel(model, x): # to extract city/region/country/location pass it␣
 ↪as X
    if x == 'city':
        return model.city
    elif x == 'region':
        return model.region
    elif x == 'country':
        return model.country
    else :
        return (model.latitude, model.longitude)
```

[14]:

```
log_df['country'] = createModel(log_df,lambda x: log_df['client_IP'])
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_7468/1480831069.py in <module>
----> 1 log_df['country'] = createModel(log_df,lambda x: log_df['client_IP'])

~\AppData\Local\Temp/ipykernel_7468/2681416502.py in createModel(db, ip)
      1 def createModel(db, ip): # create a response model holding attributes␣
 ↪and methods of location
----> 2     model = db.get(ip, api_key='Free')
      3     return model

TypeError: get() got an unexpected keyword argument 'api_key'
```

[ ]:

```
```