S04 T01: Transformació Registre Log amb Regular expressions

Descripció: L'anàlisi de registres és una funció important per al control i l'alerta, el compliment de les polítiques de seguretat, l'auditoria i el compliment normatiu, la resposta a incidents de seguretat i fins i tot les investigacions forenses. En analitzar les dades de registre, les empreses poden identificar més fàcilment les possibles amenaces i altres problemes, trobar la causa arrel i iniciar una resposta ràpida per mitigar els riscos.

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".

Primer de tot importem les llibreries que necessitem i carreguem el document amb el qual hem de treballar. El document l'he arreglat una mica amb un processador de text abans d'utilitzar-lo aquí.

In [1]:

```
import pandas as pd
import numpy as np
import re

dataf = pd.read_table("C:\\Users\\Anna\\DataScience\\SPRINTS\\SPRINT 4\\data.txt", encodir

# els punts de "encoding" i "engine" els he posat per solventar errors que em sortien a l

dataf.loc[10:30,:] #imprimeixo d'aquesta forma perquè així veig files amb localhost i amb
```

Out[1]:

	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (internal dummy connection)" VLOG=-
10	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
11	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
12	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
13	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
14	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
15	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...

	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (internal dummy connection)" VLOG=-
16	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
17	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
18	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
19	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
20	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
21	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:31 +0100	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...
22	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:33 +0100	□GET /hoteles-baratos/ofertas-hotel-Metropolis...
23	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:35 +0100	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...
24	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:38 +0100	□GET /hoteles-baratos/ofertas-hotel-Kensington...
25	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:39 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
26	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:40 +0100	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...
27	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:42 +0100	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...
28	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:45 +0100	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...
29	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:46 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
30	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:47 +0100	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...

A partir d'ara començaré a treballar sobre la base de dades "data", creant i canviant el nom de les columnes per veure més clarament els atributs amb què treballar, i també estandaritzar tota la informació

In [2]:

```
# començo per canviar els noms de les columnes que tinc per poder treballar més còmodament

dataf.columns=["pagWeb", "ip", "dataHora", "altres"]
dataf.loc[10:30,:]
```

Out[2]:

	pagWeb	ip	dataHora	altres
10	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
11	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
12	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
13	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
14	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...

	pagWeb	ip	dataHora	altres
15	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
16	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
17	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
18	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
19	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
20	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
21	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:31 +0100	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...
22	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:33 +0100	□GET /hoteles-baratos/ofertas-hotel-Metropolis...
23	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:35 +0100	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...
24	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:38 +0100	□GET /hoteles-baratos/ofertas-hotel-Kensington...
25	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:39 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
26	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:40 +0100	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...
27	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:42 +0100	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...
28	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:45 +0100	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...
29	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:46 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
30	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:47 +0100	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...

In [3]:

```
# Tipus de dades que trobem a cada columna
dataf.dtypes
```

Out[3]:

```
pagWeb      object
ip          object
dataHora    object
altres      object
dtype: object
```

In [4]:

```
# Petita descripció de cada columna
dataf.describe()
```

Out[4]:

	pagWeb	ip	dataHora	altres
count	261872	261872	261872	246399
unique	408	9276	115882	162295
top	www.akumenius.com	66.249.76.216	□GET /newdesign/libraries/anythingSlider/image...	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
freq	232300	46382	732	13891

In [5]:

```
# Com que en l'anterior punt veiem que la columna altres té menys quantitat de dades que

print(dataf.isnull())

print("_____________________________________________________________________________")

print(dataf.count())

print("_____________________________________________________________________________")
print(dataf.isnull().sum())
```

```
        pagWeb     ip  dataHora  altres
0        False  False     False   False
1        False  False     False   False
2        False  False     False   False
3        False  False     False   False
4        False  False     False   False
...        ...    ...       ...     ...
261867   False  False     False   False
261868   False  False     False   False
261869   False  False     False   False
261870   False  False     False   False
261871   False  False     False   False

[261872 rows x 4 columns]
____________________________________________________________________________
pagWeb      261872
ip          261872
dataHora    261872
altres      246399
dtype: int64
____________________________________________________________________________
pagWeb          0
ip              0
dataHora        0
altres      15473
dtype: int64
```

In [6]:

```
# Eliminem les files que contenen algun valor nul i comprobem que s'han eliminat, de mane

dataf = dataf.dropna(subset=["altres"])

dataf.describe()
```

Out[6]:

	pagWeb	ip	dataHora	altres
count	246399	246399	246399	246399
unique	2	2828	113719	162295
top	www.akumenius.com	66.249.76.216	28/Feb/2014:04:16:25 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
freq	232273	46382	83	13891

In [7]:

```
# Ara vull dividir la columna de "dataHora" en 3 columnes, la data, la de l'hora i la de

# Creo les columnes buides i les col·loco on vull i amb el nom que vull
dataf.insert(3, "data", "")
dataf.insert(4, "hora1", "")

dataf.loc[10:30,:]
```

Out[7]:

	pagWeb	ip	dataHora	data	hora1	altres
10	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
11	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...

	pagWeb	ip	dataHora	data	hora1	altres
12	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
13	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
14	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
15	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
16	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
17	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
18	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
19	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
20	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
21	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:31 +0100			□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...
22	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:33 +0100			□GET /hoteles-baratos/ofertas-hotel-Metropolis...
23	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:35 +0100			□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...
24	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:38 +0100			□GET /hoteles-baratos/ofertas-hotel-Kensington...
25	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:39 +0100			□GET /destinos-baratos/destinosEstrelles/hotel...
26	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:40 +0100			□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...
27	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:42 +0100			□GET /hoteles-baratos/ofertas-hotel-Princesa-S...
28	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:45 +0100			□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...
29	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:46 +0100			□GET /destinos-baratos/destinosEstrelles/hotel...
30	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:47 +0100			□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...

In [8]:

```
# Introdueixo la informació amb una funció lambda perquè em permeti agafar el que jo vull

dataf["data"] = dataf["dataHora"].apply(lambda x: x.split(":",1)[0])
dataf["hora1"] = dataf["dataHora"].apply(lambda x: x.split(":",1)[1])


dataf.loc[10:30,:]
```

Out[8]:

	pagWeb	ip	dataHora	data	hora1	altres
10	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
11	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
12	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
13	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
14	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
15	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
16	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
17	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
18	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
19	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
20	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
21	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...
22	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:33 +0100	23/Feb/2014	03:10:33 +0100	□GET /hoteles-baratos/ofertas-hotel-Metropolis...
23	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:35 +0100	23/Feb/2014	03:10:35 +0100	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...
24	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:38 +0100	23/Feb/2014	03:10:38 +0100	□GET /hoteles-baratos/ofertas-hotel-Kensington...
25	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:39 +0100	23/Feb/2014	03:10:39 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
26	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:40 +0100	23/Feb/2014	03:10:40 +0100	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...
27	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:42 +0100	23/Feb/2014	03:10:42 +0100	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...
28	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:45 +0100	23/Feb/2014	03:10:45 +0100	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...
29	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:46 +0100	23/Feb/2014	03:10:46 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
30	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:47 +0100	23/Feb/2014	03:10:47 +0100	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...

In [9]:

```
# Creo un nou dataset per separar l'hora i la UTC de la columna hora

hora = dataf["hora1"].str.split(" ", expand=True)
```

```
# Elimino una columna de més que se m'ha creat
hora.drop(hora.columns[2], axis=1, inplace=True)

# Canvio el nom de les columnes noves
hora.columns = ["hora", "UTC"]

#Ajunto el dataset principal "dataf" amb el dataset nou "hora"
dataf = pd.concat([dataf, hora], axis=1)

# Elimino la coumna "dataHora"
dataf = dataf.drop("dataHora", axis=1)
dataf = dataf.drop("hora1", axis=1)


dataf.loc[10:30,:]
```

Out[9]:

	pagWeb	ip	data	altres	hora	UTC
10	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
11	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
12	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
13	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
14	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
15	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
16	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
17	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
18	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
19	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
20	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
21	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...	03:10:31	+0100
22	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Metropolis...	03:10:33	+0100
23	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...	03:10:35	+0100
24	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kensington...	03:10:38	+0100
25	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:39	+0100
26	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...	03:10:40	+0100

	pagWeb	ip	data	altres	hora	UTC
27	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...	03:10:42	+0100
28	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...	03:10:45	+0100
29	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:46	+0100
30	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...	03:10:47	+0100

In [10]:

```
# Creo la columna que separi el Get i Options fins al "-" de la columna "altres"

dataf.insert(6, "optionGetComplert", "")

# Introdueixo la informació amb una funció lambda perquè em permeti agafar el que jo vull

dataf["optionGetComplert"] = dataf["altres"].map(lambda x: x.split('"-"',1)[0])

dataf.loc[10:30,:]
```

Out[10]:

	pagWeb	ip	data	altres	hora	UTC	option
10	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
11	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
12	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
13	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
14	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
15	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
16	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
17	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
18	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
19	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
20	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
21	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...	03:10:31	+0100	□GET /hoteles-ba hotel-
22	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Metropolis...	03:10:33	+0100	□GET /hoteles-ba hotel
23	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...	03:10:35	+0100	□GET /hoteles-ba hotel-

	pagWeb	ip	data	altres	hora	UTC	option
24	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kensington...	03:10:38	+0100	□GET /hoteles-ba hotel
25	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:39	+0100	□G baratos/destinosEs
26	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...	03:10:40	+0100	□GET /hoteles-ba hotel-
27	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...	03:10:42	+0100	□GET /hoteles-ba hote
28	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...	03:10:45	+0100	□GET /hoteles-ba hote
29	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:46	+0100	□G baratos/destinosEs
30	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...	03:10:47	+0100	□GET /hoteles-ba hotel

In [11]:

```
# Creo un nou dataset per separar els diferents apartats de la columna "optionGetComplert

nom = dataf["optionGetComplert"].str.split(" ", expand=True)

# Se m'han creat moltes columnes de més i les elimino utilitzant el drop
nom.drop(nom.columns[5:45], axis=1, inplace=True)

# Canvio el nom de les columnes noves
nom.columns = ["optionGet", "adreçaWeb", "HTTP", "numeroA", "numeroB"]

#Ajunto el dataset principal "dataf" amb el dataset nou "nom"
dataf = pd.concat([dataf, nom], axis=1)

# Elimino la columna "optionGetComplert" perquè ja he separat tota la informació en altres
dataf = dataf.drop("optionGetComplert", axis=1)

dataf.loc[10:30,:]
```

Out[11]:

	pagWeb	ip	data	altres	hora	UTC	optionGet
10	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
11	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
12	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
13	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
14	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
15	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
16	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
17	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS

	pagWeb	ip	data	altres	hora	UTC	optionGet	
18	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS	
19	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS	
20	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS	
21	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...	03:10:31	+0100	□GET	/hc
22	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Metropolis...	03:10:33	+0100	□GET	/hc
23	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...	03:10:35	+0100	□GET	/hc
24	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kensington...	03:10:38	+0100	□GET	/hc
25	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:39	+0100	□GET	barat
26	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...	03:10:40	+0100	□GET	/hc
27	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...	03:10:42	+0100	□GET	/hc
28	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...	03:10:45	+0100	□GET	/hc
29	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:46	+0100	□GET	barat
30	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...	03:10:47	+0100	□GET	/hc

In [12]:

```python
#Repeteixo el procés anterior per a l'última part de la columna altres

# Creo un nou dataset per separar l'última part de la columna altres
final = dataf["altres"].str.split('"-"', expand=True)

# Em quedo només amb la columna que vull treballar
final.drop(final.columns[0], axis=1, inplace=True)
final.drop(final.columns[1], axis=1, inplace=True)

# Canvio el nom de les columnes noves
final.columns = ["colFinal"]

print(final)
```

```
                                                 colFinal
0              "Apache (internal dummy connection)" VLOG=-
1              "Apache (internal dummy connection)" VLOG=-
2              "Apache (internal dummy connection)" VLOG=-
3              "Apache (internal dummy connection)" VLOG=-
4              "Apache (internal dummy connection)" VLOG=-
...                                                   ...
261867   "Mozilla/5.0 (compatible; YandexBot/3.0; +htt...
261868   "Mozilla/5.0+(compatible; UptimeRobot/2.0; ht...
261869         "Apache (internal dummy connection)" VLOG=-
```

```
261870         "Apache (internal dummy connection)" VLOG=-
261871         "Apache (internal dummy connection)" VLOG=-

[246399 rows x 1 columns]
```

In [13]:

```
#Ajunto el dataset principal "dataf" amb el dataset nou "final"
dataf = pd.concat([dataf, final], axis=1)

# Elimino la columna "altres" perquè ja he separat tota la informació en altres columnes
dataf = dataf.drop("altres", axis=1)

dataf.loc[10:30,:]
```

Out[13]:

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb	
10	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
11	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
12	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
13	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
14	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
15	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
16	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
17	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb	
18	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
19	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
20	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
21	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:31	+0100	□GET	/hoteles-baratos/ofertas-hotel-Club-&-Hotel-Le...	HTT
22	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:33	+0100	□GET	/hoteles-baratos/ofertas-hotel-Metropolis-Hote...	HTT
23	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:35	+0100	□GET	/hoteles-baratos/ofertas-hotel-Faena-Hotel-Bue...	HTT
24	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:38	+0100	□GET	/hoteles-baratos/ofertas-hotel-Kensington-Town...	HTT
25	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:39	+0100	□GET	/destinos-baratos/destinosEstrelles/hoteles-en...	HTT
26	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:40	+0100	□GET	/hoteles-baratos/ofertas-hotel-Howard-Jhonson-...	HTT
27	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:42	+0100	□GET	/hoteles-baratos/ofertas-hotel-Princesa-Sofia-...	HTT
28	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:45	+0100	□GET	/hoteles-baratos/ofertas-hotel-Kfar-Giladi-en-...	HTT
29	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:46	+0100	□GET	/destinos-baratos/destinosEstrelles/hoteles-en...	HTT

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb	
30	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:47	+0100	□GET	/hoteles-baratos/ofertas-hotel-Casual-Hotel-Va...	HTT

In [14]:

```
# Separo en diferents columnes la informació de la columna restant
final1 = dataf["colFinal"].str.split('"', expand=True)

# Canvio el nom de les columnes noves
final1.columns = ["buscador", "apache", "vlog"]

# Elimino la columna "colFinal" perquè ja he separat tota la informació en altres columnes
dataf = dataf.drop("colFinal", axis=1)

#Ajunto el dataset principal "dataf" amb el dataset nou "final1"
dataf = pd.concat([dataf, final1], axis=1)



dataf.loc[10:30,:]
```

Out[14]:

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb	
10	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
11	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
12	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
13	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
14	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
15	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
16	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
17	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb	
18	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
19	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
20	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
21	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:31	+0100	□GET	/hoteles-baratos/ofertas-hotel-Club-&-Hotel-Le…	HTT
22	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:33	+0100	□GET	/hoteles-baratos/ofertas-hotel-Metropolis-Hote…	HTT
23	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:35	+0100	□GET	/hoteles-baratos/ofertas-hotel-Faena-Hotel-Bue…	HTT
24	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:38	+0100	□GET	/hoteles-baratos/ofertas-hotel-Kensington-Town…	HTT
25	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:39	+0100	□GET	/destinos-baratos/destinosEstrelles/hoteles-en…	HTT
26	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:40	+0100	□GET	/hoteles-baratos/ofertas-hotel-Howard-Jhonson-…	HTT
27	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:42	+0100	□GET	/hoteles-baratos/ofertas-hotel-Princesa-Sofia-…	HTT
28	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:45	+0100	□GET	/hoteles-baratos/ofertas-hotel-Kfar-Giladi-en-…	HTT
29	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:46	+0100	□GET	/destinos-baratos/destinosEstrelles/hoteles-en…	HTT
30	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:47	+0100	□GET	/hoteles-baratos/ofertas-hotel-Casual-Hotel-Va…	HTT

NIVELL 2

Exercici 2

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

```
In [ ]:
```

Exercici 3

Geolocalitza les IP's.

```
In [ ]:
```

NIVELL 3

Exercici 4

Mostra'm la teva creativitat, Sorprèn-me fes un pas més enllà amb l'anàlisi anterior.