

Entrega Actividad: **Generación de informes**

Analítica Web(AW)

Máster Universitario en Inteligencia Computacional e Internet de las Cosas

Universidad de Córdoba, EPSC

2022/2023



UNIVERSIDAD
DE
CÓRDOBA

Autor:

Antonio Gómez Giménez (i72gogia@uco.es)

En este ejercicio se presenta el Dataset Dataiku, que contiene información sobre cada vista de página individual en el sitio web www.Dataiku.com. (información de acceso durante dos meses).

La estructura de dicho dataset es la siguiente:

- **server_ts**: fecha de conexión al servidor
- **client_ts**: fecha de conexión del usuario
- **client_addr**: dirección IP del usuario
- **visitor_id**: id asignado al usuario (~ _utma)
- **session_id**: id de la sesión/visita (~ _utmb)
- **location**: url de la página a la que se accede
- **referrer**: url de la página desde la que se accedió a location. Si está vacío se considera una búsqueda directa
- **user_agent**: navegador desde el que se conectó el usuario
- **type**: tipo de acceso
- **visitor_param**:
- **sesion_param**:
- **event_param**:
- **br_width**: ancho del navegador del usuario
- **br_height**: alto del navegador del usuario
- **sc_width**: ancho de la pantalla del usuario (resolución)
- **sc_height**: alto de la pantalla del usuario (resolución)
- **br_lang**: idioma del navegador del usuario
- **tz_off**: time zona. El número indica la diferencia en minutos con la hora GMT . -60 indica zona horaria GMT+1 y 60 indica zona horaria GMT-1.

Teniendo esto en cuenta, el ejercicio como tal, consiste en, a partir del dataset anterior, generar informes para, en base a dichos datos, poder tomar las siguientes decisiones:

- Qué idiomas debería soportar el sitio web.
- Se quiere sacar una promoción durante un tiempo muy limitado. Qué hora sería la más adecuada para llegar a un mayor número de usuarios.
- Podemos decidir promocionarse en los enlaces que vienen de serie con los navegadores. En qué navegador sería más interesante hacernos publicidad.

Para poder llevar a cabo dicho ejercicio, se usó el lenguaje de programación python, donde primeramente se cargó dicho dataset.

El código es el siguiente:

```
1 import pandas as pd
2
3 Dataset = pd.read_csv("LogsDataiku.csv", sep=',')
```

Primeramente, para responder a la primera decisión (Qué idiomas debería soportar el sitio web), es necesario contar la frecuencia de cada idioma de navegador que ha usado cada usuario único. Para ello se creó el siguiente código:

```
5 Dataset = Dataset.groupby(by="visitor_id").nth(0)#obtengo para cada usuario su primera acceso al entrar al sitio web
6 print((Dataset['br_lang'].value_counts()))#usar value_count para agrupar por lenguaje
```

El resultado obtenido es el siguiente:

fr	971
en-US	846
en-us	183
fr-fr	138
en-GB	71

Por lo tanto, los idiomas principales con los que se debería trabajar en el sitio web son el francés y el inglés.

Para responder a la segunda decisión (Se quiere sacar una promoción durante un tiempo muy limitado. Qué hora sería la más adecuada para llegar a un mayor número de usuarios), es necesario contar la frecuencia de usuarios únicos respecto a las diferentes horas del día, para saber cuál es la hora más transitada. Para ello se creó el siguiente código:

```
8 Dataset['Datetime'] = pd.to_datetime(Dataset['server_ts'])
9 Dataset['hour'] = pd.to_datetime(Dataset['Datetime']).dt.strftime("%H"+" :00")
10 print(Dataset['hour'].value_counts())
```

El resultado es el siguiente:

15:00	265
14:00	237
16:00	179
13:00	170
09:00	137
12:00	137
17:00	136
11:00	132
10:00	120
20:00	112
21:00	107
19:00	106
08:00	106
22:00	95
18:00	80
07:00	73
06:00	55
02:00	52
23:00	48
01:00	44
03:00	41
00:00	41
05:00	38
04:00	26

Como se puede apreciar en la imagen anterior, la hora más transitada es a las 3 de la tarde, por tanto se recomienda realizar la promoción durante esta hora.

Para responder a la tercera decisión (Podemos decidir promocionarse en los enlaces que vienen de serie con los navegadores. En qué navegador sería más interesante hacernos publicidad), para ello sería necesario calcular la frecuencia de cada navegador usado por cada usuario único. Para ello se creó el siguiente código:

```
29 Dataset = Dataset.groupby(by="visitor_id").nth(0)#obtengo para cada usuario su primera acceso al entrar al sitio web
30 print((Dataset['user_agent'].value_counts()))#usar value_count para agrupar por lenguaje
```

El resultado es el siguiente:

```
Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1750.154 Safari/537.36 228
Mozilla/5.0 (Windows NT 6.1; WOW64; rv:27.0) Gecko/20100101 Firefox/27.0 156
Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1750.146 Safari/537.36 138
Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_2) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1750.152 Safari/537.36 109
Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_2) AppleWebKit/537.74.9 (KHTML, like Gecko) Version/7.0.2 Safari/537.74.9 66
...
Mozilla/5.0 (Linux; U; Android 4.2.2; en-us; GT-P5113 Build/JDQ39) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Safari/534.30 1
Mozilla/5.0 (compatible; MSIE 10.0; Windows Phone 8.0; Trident/6.0; IEMobile/10.0; ARM; Touch; NOKIA; Lumia 920) 1
Mozilla/5.0 (Windows NT 5.1; rv:24.0) Gecko/20100101 Firefox/24.0 1
Mozilla/5.0 (hp-tablet; Linux; hpwOS/3.0.5; U; en-US) AppleWebKit/534.6 (KHTML, like Gecko) wOSBrowser/234.83 Safari/534.6 TouchPad/1.0 1
Mozilla/5.0 (Windows NT 6.3; WOW64; Trident/7.0; Touch; MDDCJS; rv:11.0) like Gecko 1
```

Como podemos observar, el navegador más usado es mozilla, por tanto, se recomienda que si se desea realizar una campaña sobre algún tipo de navegador, sea sobre mozilla (Firefox).