Entrega Actividad: Cálculo de métricas de un sitio web

Analítica Web(AW)

Máster Universitario en Inteligencia Computacional e Internet de las Cosas

Universidad de Córdoba, EPSC

2022/2023

UNIVERSIDAD
D
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
- **referer**: 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_lan**: 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, calcular las siguientes métricas para el periodo de tiempo registrado:

- **N.º de visitas** → Número de visitas (sesiones que ha tenido el sitio web).
- **N.º de visitantes únicos** → Número de usuarios diferentes que han visitado el sitio web.
- **N.º medio de páginas/visitas** →Para cada visita (sesión) cuántas páginas se han visitado. Media para todas las visitas
- **Tasa de rebote** → Número de visitas (sesiones) que solo han accedido a una página.
- **Tasa de salida para cada página** → % de veces que cada página ha sido una página de salida.
- **Tráfico directo** → Número de visitas (sesiones) que provienen de escribir la url directamente.

- **Tráfico de búsqueda** → Número de visitas (sesiones) que provienen de pinchar en una búsqueda.
- **Tráfico referido** → Número de visitas (sesiones) que provienen de pinchar en un enlace de otra página.

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 calcular el **número de visitas** que se han obtenido en el sitio web, es necesario contar las sesiones únicas realizadas. Para ello se creó el siguiente código:

```
5   print(len(Dataset['session_id'].unique()))
```

El resultado obtenido es un total de 3946 visitas.

Para poder comprobar el número de **visitantes únicos**, se calculará el total de id de visitantes únicos, para así obtener este parámetro. El código realizado es el siguiente:

```
7   print(len(Dataset['visitor_id'].unique()))
```

El resultado obtenido es un total de 2538 visitantes únicos.

Para poder comprobar el **número medio de páginas/visitas**, para cada visita (sesión) se comprobará cuántas páginas se han visitado y se realizará la media para todas las visitas. El código realizado es el siguiente:

```
9    print(Dataset['session_id'].value_counts().mean())
```

En dicho código, se está calculando la frecuencia de cada sesión(páginas visitadas) y se realiza la media para sacar el número de páginas/visitas para cada sesión.

El resultado obtenido es de una media de 2.749 páginas por sesión.

Para calcular la **tasa de rebote**, se calculó el número de visitas (sesiones) que solo han accedido a una página. Para ello, se realizó el siguiente código:

```
11    print((Dataset['session_id'].value_counts() == 1).sum())
```

Como se puede observar, agrupamos por id y del conjunto escogemos aquellos valores que están a 1 (indica el rebote ya que en una sesión solo accede a una página)

El resultado obtenido es de una tasa de 2165 de rebote.

Para calcular la **tasa de salida para cada página**, se calculará el porcentaje de veces que cada página ha sido una página de salida. Para ello, primero se obtuvo la frecuencia de aparición de cada página (en total hay 96, teniendo cada una su respectiva frecuencia de aparición) y se calculó también, la frecuencia de aparición de cada página siendo esta la última de una sesión(en total hay 76, teniendo en cada una su respectiva frecuencia de aparición). Si se quisiera comprobar para cada página, simplemente habría que dividir la frecuencia de cada página que ha sido última sesión entre la frecuencia de dicha página. También se podría realizar como la frecuencia de salida de una página entre la cantidad de sesiones, ya que para cada sesión hay una salida.
Como son demasiadas páginas (en concreto 76 páginas que son de salida), para simplificar, se realizará la frecuencia de páginas de salida totales, entre la frecuencia de páginas. El código usado es el siguiente:

```
12
13   print(Dataset['location'].value_counts())#numero total de veces que se repite cada página
14   Dataset_tmp = Dataset.loc[:, ['session_id', 'location']]#escojo columnas que me interesan
15   Dataset_tmp = Dataset_tmp.groupby(by="session_id").nth(-1)#obtengo para cada session la última página
16   print((Dataset_tmp['location'].value_counts()))#numero total de veces que se repite cada página siendo la última
17   print(Dataset_tmp['location'].value_counts().sum()/Dataset['location'].value_counts().sum())#tasa de salidas frente a numero de páginas
18
19
```

El resultado obtenido es una tasa de 0.3637.

Para obtener el **tráfico directo**, calcularemos el número de visitas (sesiones) que provienen de escribir la url directamente. Para ello comprobaremos cuales tienen el campo referer vacío. El código es el siguiente:

```
21
22    print(pd.isna(Dataset['referer']).sum())
23
```

El resultado es 1226 de tráfico directo.

Para obtener el **tráfico de búsqueda**, calcularemos el número de visitas (sesiones) que provienen de pinchar en una búsqueda. Para ello comprobaremos cuales no tienen el campo referer vacío y coinciden con la cadena "www.dataiku.com/blog/", ya que las búsquedas se realizan en dicha página. El código es el siguiente:

```
23
24    Dataset_tmp = Dataset.loc[:, ['session_id', 'referer']]#escojo columnas que me interesan
25    Dataset_tmp = Dataset_tmp.groupby(by="session_id").nth(0)#obtengo para cada session la primera páginaS
26    Dataset_tmp=Dataset_tmp.dropna(axis=0)#elimino filas con valores vacios ya que son por tráfico directo
27    print(Dataset_tmp.loc[Dataset_tmp['referer'].str.contains('www.dataiku.com/blog/')==True].count())#escojo aquellos que son por busqueda
```

El resultado es un total de 51.

Para obtener el **tráfico de búsqueda**, calcularemos el número de visitas (sesiones) que provienen de pinchar en una búsqueda. Para ello comprobaremos cuales no tienen el campo referer vacío y no coinciden con la cadena "www.dataiku.com/blog/", ya que los accesos serán por cualquier tipo de página externa. El código es el siguiente:

```
28
29
30    print(Dataset_tmp.loc[Dataset_tmp['referer'].str.contains('www.dataiku.com/blog/')==False].count())#escojo aquellos que son por link de otra página
31
32
33
```

El resultado es un total de 2871.