PRAC1

WEB-SCRAPPING

EVGENY MUZAREV GEVORGIAN – PABLO CHILLERÓN BEVIÁ

NOVIEMBRE 2021

PRAC1 – TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Contenido

Ejercicio 1 - Contexto

Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web proporciona dicha información.

En el contexto del supuesto práctico propuesto en esta práctica, necesitamos un portal de internet que contenga datos y precios sobre los precios de habitaciones de hoteles de diferentes lugares, características y fechas.

Booking.com es el portal más usado por establecimientos para comercializar con sus servicios, operando en más de 200 países y casi 80.000 destinos diferentes. Por otro lado, más de 40 millones de usuarios que consultan el portal antes de reservar sus vacaciones.

Es, por tanto, un sitio web excelente para extraer la información que necesitamos.

Ejercicio 2 - Título

Definir un título que sea descriptivo para el dataset.

“Situación del precio de la habitación en la costa española en los próximos días”.

Ejercicio 3 – Descripción del dataset

Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

El dataset está dividido en dos partes:

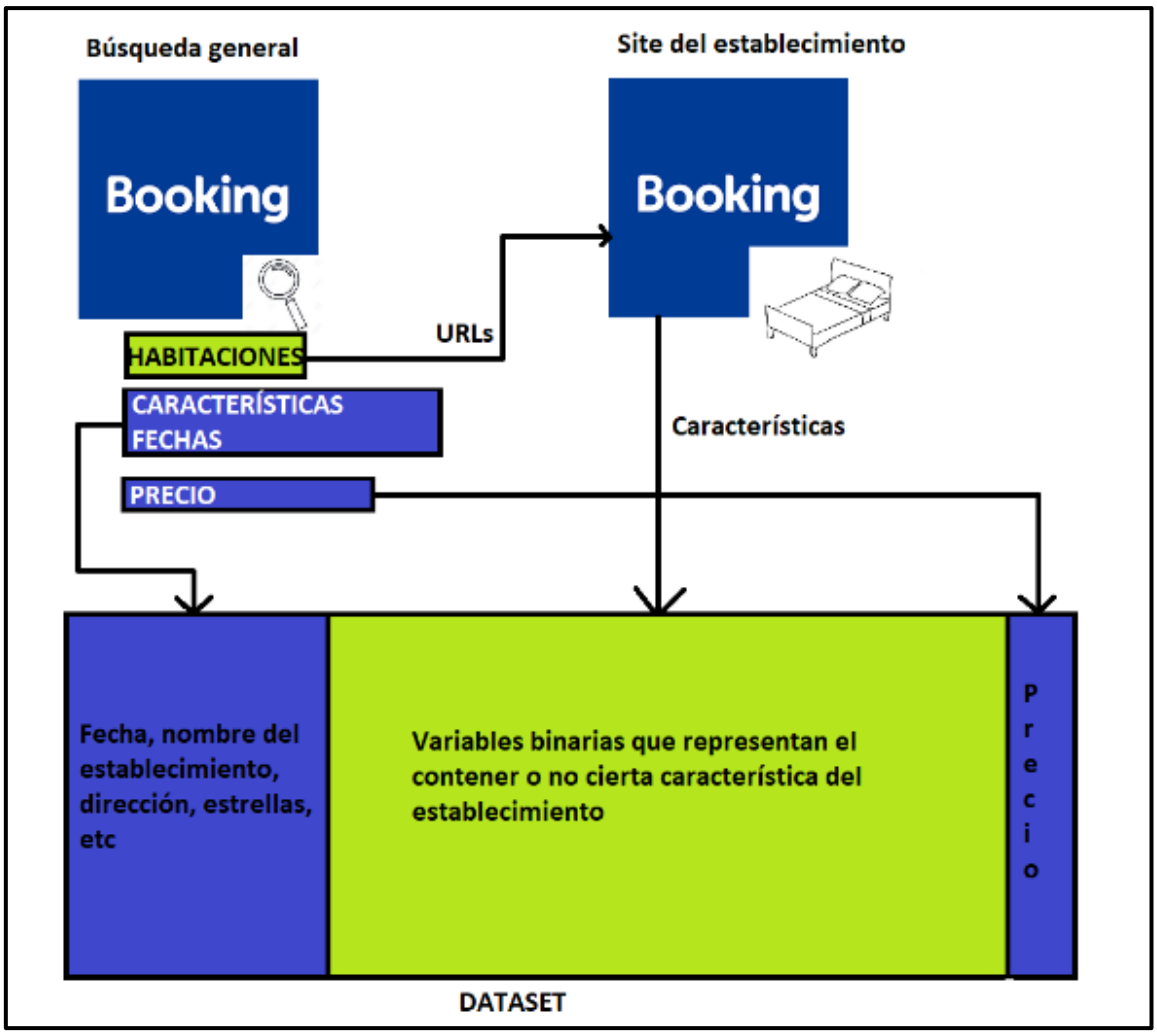
a) Una primera parte que, para cada fecha, describe sus dimensiones de identidad, geográficas y otras características relacionadas con el establecimiento, su ubicación y el precio de la habitación para esa noche.
b) Otras 26 características como el tener restaurante, ascensor o piscina entre otras, que nos puede ayudar para hacer diversos análisis, con el fin de poder relacionar estas características con el precio de la noche.

La unión de estas dos partes, crea el dataset completo, donde encontramos para cada fecha y establecimiento, una descripción del mismo a través de ciertas características y el precio de la noche.

Los establecimientos del dataset, están previamente filtrados, por lo que no es necesario ningún filtro interno para obtener aquellos establecimientos situados en la costa española.

Ejercicio 4 – Representación gráfica

Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



Ejercicio 5 - Contenido

Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

A continuación, se exponen los campos que incluye el dataset, así como su tipo y una breve descripción:

<u>Atributos sobre la identidad del establecimiento:</u>

Id (Entero) – Atributo autonumérico que enumera cada registro.

Date (Fecha) – Fecha que representa la noche de la reserva

Name (Texto) – Nombre del establecimiento

Stars (Entero) – Número de estrellas del establecimiento

Address (Texto) – Dirección del establecimiento

Country (Texto) – País del establecimiento

City (Texto) – Ciudad del establecimiento

Postal_Code (Entero) – Código postal del establecimiento

Score (Decimal) – Puntuación media de los clientes al establecimiento

Comments (Entero) – Número de comentarios que tiene el establecimiento en el portal

Beach (Decimal) – Distancia en kilómetros del establecimiento a la playa

Price (Entero) – Precio de la noche para la fecha del registro.

Atributos sobre las características del establecimiento:

Todas estas variables, son binarias, donde:

1 significa que contiene esa característica o servicio

0 significa que no contiene esa característica o servicio

Airport Shuttle (free) – Lanzadera al aeropuerto gratuita

BBQ facilities – Espacio para hacer barbacoas

Free WiFi Internet Access Included – Conexión WIFI gratuita

Wireless Lan – Conexión Wireless LAN incluida

Daily maid service – Servicio de limpieza de habitaciones diario

Swimming pool - Piscina

Pets allowed – Las mascotas están permitidas

Free Parking – Parking gratuito

Spa & Wellness Centre – Servicio de SPA y bienestar

Fitness Room – Sala fitness

Private Beach Area – Playa privada

Heating - Calefacción

Family Rooms – Habitaciones familiares

Coffee/Tea maker – Máquina de café/té

Restaurant - Restaurante

Beach front – A pie de playa

Garden - Jardín

Bar - Bar

Terrace - Terraza

Rooms/Facilities for Disabled – Habitaciones adecuadas para personas con movilidad reducida

Elevator - Ascensor

Non Smoking Rooms – Habitaciones de no fumadores

24 hour Front Desk – Recepción 24 horas

Room-service – Servicio de habitaciones

Airport Shuttle – Lanzadera al aeropuerto (no gratuito)

Parking (fee required) – Parking (no gratuito)

En cuanto al periodo de tiempo de los datos, el dataset publicado contiene registros fijos de 14 días (22/10/2021 – 04/11/2021), sin embargo, el script publicado en GitHub, genera un nuevo dataset con registros desde la fecha de lanzamiento del script hasta los siguientes 14 días.

El dataset se ha recogido mediante web-scrapping, usando un script de Python publicado en GitHub.

Ejercicio 6 - Agradecimientos

Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Existen múltiples datasets relacionados con el sector de la hostelería, principalmente, desde el punto de vista del cliente, como, por ejemplo, qué época es la mejor para viajar, pero no se hemos encontrado ninguno desde el punto de vista del establecimiento. Existen varias plataformas de pago para los establecimientos, que, mediante acceso a sus ERPs, les muestra el precio de la noche para su competencia directa, definiendo competencia como aquellos establecimientos similares en cuanto a localización, características y servicios.

Cabe destacar, que existen varios ejemplos de web-scrapping basados en booking.com, aunque en este estudio vamos más allá, poniéndonos en la mente del consumidor que no tiene claro dónde quiere ir de vacaciones, pero sí el tipo de establecimiento que busca.

En este aspecto, booking.com contiene multitud de datos de establecimientos de todo el mundo que permite analizar al cliente y comparar el establecimiento con la competencia.

Con el fin de limitarnos en todo momento de acuerdo con los principios legales y éticos, hemos seguido los siguientes pasos:

1) Analizar la API de booking.com y comprobar que no se adapta a nuestro contexto.
2) Consultar el archivo robots.txt de booking.com (https://www.booking.com/robots.txt), donde podemos comprobar que no existe ninguna restricción en nuestro estudio.

Por otro lado, con el fin de esquivar posibles obstáculos en este proceso, se ha tenido en cuenta lo siguiente:

1) Modificar el user-agent de los headers.
2) Utilizar la función time.sleep() para no saturar las peticiones al servidor web.

Ejercicio 7 - Inspiración

Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

El dataset obtenido tiene muchas aplicaciones para los establecimientos, tal y como se puede comprobar en el script.

Entre otras cosas, este dataset es interesantes porque:

1) Se puede analizar a la competencia desde el punto de vista del consumidor.
2) Permite comprobar si los precios son realmente atractivos para el consumidor, en comparación con establecimientos similares.
3) Visualiza la oferta existente a nivel nacional, lo que permite conocer los puntos fuertes y débiles del establecimiento.

Por otro lado, estos son algunos estudios posibles:

1) ¿Existe alguna manera diferente de agrupar los hoteles obviando el número de estrellas? Análisis cluster.
2) ¿Es posible predecir el precio de una habitación dadas ciertas características? Creación de presupuestos usando modelos de regresión.
3) ¿Qué características de un establecimiento son más importantes para predecir el precio de una noche? Predicciones sobre intervalos de precios mediante árboles de decisión.
4) ¿Existe alguna agrupación entre características de los hoteles que tengan alguna relevancia? Algoritmos de agrupación.

Este tipo de estudios entroncan perfectamente con el objetivo de analizar los establecimientos desde el punto del vista del consumidor, ya que, recordemos, el objetivo es ponerse en lugar del consumidor sin un rumbo determinado, pero sí con un tipo de establecimiento y unas características determinadas.

Ejercicio 8 - Licencia

Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:

- **Released Under CC0: Public Domain License.**
- **Released Under CC BY-NC-SA 4.0 License.**
- **Released Under CC BY-SA 4.0 License.**
- **Database released under Open Database License, individual contents under Database Contents License.**
- **Other (specified above).**
- **Unknown License.**

La licencia seleccionada es RELEASED UNDER CC BY-NC-SA 4.0 LICENCE, ya que este dataset ha sido creado como ejemplo de potencial del web-scrapping y de sus múltiples aplicaciones.

El usuario puede hacer lo siguiente:

1) Copiar y redistribuir el dataset, ya sea en otro medio o formato.
2) Transformar el dataset y adaptarlo a nuevas necesidades de estudio.
3) Citar un enlace al dataset original en caso de ser usado, así como indicar los cambios realizados en el mismo.
4) Distribuir el dataset resultante tras transformaciones bajo la misma licencia.
5) A pesar de la naturaleza y contexto de este dataset, no usarlo con propósitos comerciales.

Ejercicio 9 - Código

Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

Link: https://github.com/evgmg/web_scraping

Ejercicio 10 - Dataset

Publicar el dataset obtenido en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

```
DOI Link: 10.5281/zenodo.5639715
Target URL: https://doi.org/10.5281/zenodo.5639715
```

Contribuciones

Contribuciones	Firmas
Investigación previa	Evgeny Murarez Gevorgian, Pablo Chillerón Beviá
Redacción de las respuestas	Evgeny Murarez Gevorgian, Pablo Chillerón Beviá
Desarrollo del código	Evgeny Murarez Gevorgian, Pablo Chillerón Beviá