Tipología y ciclo de vida de los datos - PRÁCTICA 1

Aysha Ait Ouaddi El Mamouny
Oriol González Dalmau

1. Contexto

España es desde hace unos años un ícono del Turismo internacional, este año han visitado el territorio español hasta agosto 48 millones de turistas extranjeros [1], después de la caída por la pandemia de los visitantes, los números se han recuperado quedando cerca de los niveles de 2019 [2], así pues, el turismo es uno de los principales motores económicos del país, que según estimaciones podría haber aportado hasta un 14% del PIB de la economía española en el 2019 [3].

El turismo se mueve por regiones que a priori son desconocidas, cada visitante tendrá un motivo de visita, un periodo de estancia y un rango kilométrico predefinido. Al escoger una destinación, los turistas se ven condicionados por el motivo de su viaje, pero también por las plazas de hospedaje disponibles, lo que acaba repercutiendo en el coste de la pernoctación.

Entender el flujo de personas de una ciudad podría influir en la capacidad de prever otros factores que afectan al estado de un entorno dinámico como es una ciudad ¿Qué información hay que nos indique cómo de llena está una ciudad? ¿Cómo se puede saber de donde es el turista que nos visita? Una de las principales limitaciones a la hora de obtener datos, es que pese a que los hoteles publican las estadísticas de ocupación que han obtenido, no lo hacen hasta que ha llegado el final del mes. Además, las estadísticas muestran los datos agregados y no se puede conocer la ocupación detallada, ni mucho menos el precio de venta por habitación.

Reservar unas vacaciones es ahora más fácil que nunca. Con un búsqueda a través de un *smartphone* se puede organizar un viaje. Un ejemplo de una plataforma que lo permite es eDreams, https://www.edreams.com/.

Se ha escogido eDreams para la realización del proyecto dado que dispone de muchos datos y métricas siendo la plataforma de referencia para este tipo de negocio. En esta se pueden encontrar todos los servicios necesarios para un viaje al completo, lo que como usuarios permite acceder a variedad de servicios y como científicos de datos posibilita recolectar esta información mediante web scraping, con el objetivo de aplicar modelos de aprendizaje y poder obtener información estadísticamente significativa de forma alternativa a las encuestas de ocupación. El siguiente enlace [4] muestra una búsqueda de hoteles para la ciudad de Barcelona.

2. Título

El título que representa el conjunto de datos obtenido es:
“Hoteles de Barcelona disponibles el día 30 de noviembre: los datos más relevantes.”.

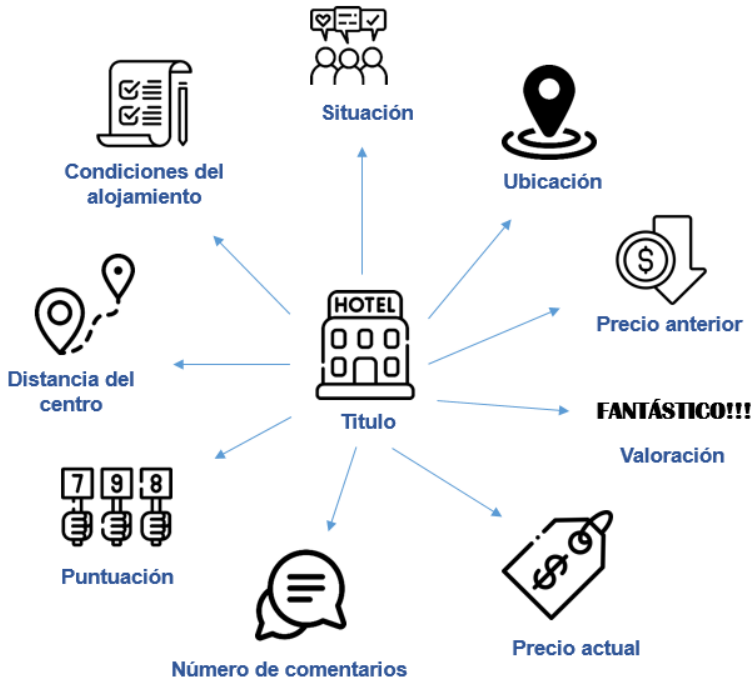
3. Descripción del dataset

El dataset recoge una muestra de los hoteles disponibles en Barcelona el día 30 noviembre, previamente extraídos de la plataforma eDreams. Asimismo, también recoge datos de importancia referentes a los hoteles, como pueden ser la distancia al centro de la ciudad, los comentarios recibidos, la valoración o el precio, entre otros. Por lo tanto, se trata de un conjunto de datos que ofrece posibilidades de explotación analítica.

Por otro lado, cabe resaltar el hecho de que los datos no han sido procesados previamente, por lo que pueden realizarse tareas de limpieza que permitan disponer de una mejor calidad, dado que existen variables almacenadas como tipo string, las cuales pueden ser manipuladas con el objetivo de convertirlas en tipo numéricas.

Por último, el dataset se recoge en un fichero CSV con el fin de facilitar su manipulación y almacenamiento.

4. Representación gráfica



5. Contenido

El periodo de tiempo es de carácter diario dado que los datos hacen referencia al día 30 de noviembre. La estructura del dataset está conformada por 10 variables y 75 registros.

- Título: Nombre oficial del hotel. Variable tipo string.
  Ejemplo: Rooms Balmes
- Ubicación: Distrito en el que se encuentra ubicado el hotel. Variable tipo string.
  Ejemplo: Eixample, Barcelona
- Valoración: Valoración del hotel. Variable tipo string.
  Ejemplo: Fantástico
- Número de comentarios: Número de comentarios que han sido realizados por clientes. Variable tipo string
  Ejemplo: 2702 comentarios
- Puntuación: Nota recibida. Variable numérica
  Ejemplo: 8,3
- Distancia del centro: Distancia aproximada que separa el hotel del centro de la ciudad. Variable tipo string
  Ejemplo: 2,2 km del centro.

- Condiciones del alojamiento: Características de la habitación. Variable tipo string.
  Ejemplo: Habitación doble.
- Situación: Valoración de la ubicación en la que se encuentra el hotel. Variable tipo string.
  Ejemplo: Ubicación 9,4
- Precio anterior: Precio anterior del servicio hotelero (sin descontar). Variable tipo string.
  Ejemplo: € 129
- Precio actual: Precio actual del servicio hotelero (descontado). Variable tipo string.
  Ejemplo: € 79

Los datos fueron recopilados mediante un proceso de web scraping en Python, sobre las cuatro primeras páginas proporcionadas por la plataforma eDreams tras realizar una búsqueda de un día concreto (30 de noviembre).

Finalmente, es importante destacar que debido a las circunstancias de la plataforma, existe la posibilidad de que los datos referentes a los hoteles en la búsqueda concreta se actualicen y varíe con respecto al contenido del dataset original. Un ejemplo puede ser debido a la variación del precio del servicio, la indisponibilidad de un hotel o un aumento en el número de comentarios.

6. Propietario

Los datos del dataset pertenecen a la página web de la agencia de viajes online eDreams, esta fue fundada en Barcelona por Javier Perez-Tenessa y tiene como objetivo ofrecer precios económicos en vuelos, hoteles, alquileres de coches y paquetes vacacionales entre otros.

El objeto de estudio han sido los hoteles de la ciudad de Barcelona, a su vez eDreams ha obtenido los precios mediante los Global Distribution Systems (GDS) como Amadeus, Galileo o Sabre (aunque luego se harán ofertas o se sumará su comisión dependiendo del contrato que tengan con las compañías hoteleras).

Antes de empezar con un proyecto de web scraping es preciso verificar que es posible acceder a los datos de forma recurrente con un bot. Es por ello que revisar el archivo robots.txt es una buena praxis, ya que indica cómo de accesible son sus páginas. Se observa una captura del mismo:

edreams.com/robots.txt

```
User-Agent: *
Disallow: /cgi-bin/        # Disallowed for obvious reasons...
Disallow: /*sessionid
Disallow: /flights-offers/
Allow: /travel/setup.js/index.jsp?noext=1
Disallow: /travel/
Disallow: /*mktportal
Disallow: /*adlabel
Disallow: /*zanpid
Disallow: /marketing-channel/v1/track

User-agent: AdsBot-Google-Mobile
Disallow:

User-agent: AdsBot-Google
Disallow:

User-Agent: AdNetTrack
Disallow: /

User-Agent: SearchTone2.0
Disallow: /

User-Agent: LEIAcrawler
Disallow: /

User-Agent: WhatsUp_Gold/5.01
Disallow: /

User-Agent: MindCrawler
Disallow: /

# PS-414
User-agent: Mediapartners-Google
Disallow:
```

Como se puede observar, no especifica que no se pueda realizar un proceso de web scraping a la información de los hoteles. Además, es interesante remarcar que según los términos y condiciones de la página web, concretamente en los derechos de propiedad intelectual e industrial, se indica que todo el contenido de esta plataforma (incluido, a título informativo y no limitativo, marcas comerciales, textos, gráficos, logotipos, iconos de botones, imágenes, archivos de audio y software) es propiedad de eDreams o de sus proveedores de contenido y está protegido por las leyes nacionales e internacionales de propiedad intelectual e industrial. [5] Asimismo, dicho documento tampoco hace referencia a la oposición de realizar web scraping sobre la plataforma.

7. Inspiración

Así como lo hacen otras industrias, el turismo es una actividad económica que genera muchos datos que pueden ser de relevante interés para los agentes que participan en dicha actividad. Por lo tanto, el conjunto de datos que se ha recogido permite responder a múltiples preguntas. Algunas de ellas son:

- Distritos de Barcelona que proporcionan una mayor oferta hotelera.
- Ranking de hoteles con mejores valoraciones.
- Determinar la correlación entre variables.
- Valoración media por hoteles de cada distrito.
- Clustering de los hoteles por precio.
- Desarrollar modelos de predicción de precios o puntuación.

El conjunto de datos es abierto al dominio público y no encontramos limitaciones para la automatización de la información, así pues se nos permite trabajar sobre los datos obtenidos y no hay límites externos a lo que podamos obtener de los datos.

Sin embargo, debemos tener presente que una de las principales limitaciones de este dataset es que los datos varían con mucha frecuencia, incluso varias veces por día, por lo que la descarga de información debe ser monitoreada constantemente y aspirar a obtener reportes directos vinculados con la descarga automática para actualizar este dataset cada vez que sea necesario.

8. Licencia

Tras la evaluación de las diferentes licencias propuestas, se determina que la licencia óptima para este proyecto es la Attribution-ShareAlike 4.0 International (o Released Under CC BY-SA 4.0 License)[6].

La decisión se fundamenta en que permite que otros utilicen, modifiquen y construyan a partir del trabajo publicado, incluso con fines comerciales, siempre que le den crédito y licencien sus nuevas creaciones bajo los mismos términos. Esto es debido a que se desea que los usuarios finales puedan construir casos de uso a partir de los datos ofrecidos. Además, también se caracteriza por la no responsabilidad del material licenciado y la exención de responsabilidad sobre el material ofrecido, quedando exentos de manera jurídica.

Por ejemplo, si los datos no fuesen consistentes y terceros obtienen conclusiones relevantes, esta licencia exime de responder ante erratas producidas por la falta de consistencia dicha.

9. Código

Puede consultarse el código utilizado dentro de la carpeta source del repositorio "Código scraping hoteles eDreams .py" :
https://github.com/Ayshaait/Scraping-de-hoteles-en-la-plataforma-eDreams

El código Python diseñado para este proyecto consta de dos partes diferenciadas. Por un lado, una primera fase en que se consulta el archivo "robots.txt" de la plataforma eDreams, se comprueba la tecnología usada y se determina el propietario de la página web. Por otro lado, el código que permite realizar la tarea de web scraping. Esta consiste en una fase de importación de librerías necesarias, el establecimiento del user agent, la inicialización del driver Chrome, el acceso al enlace objetivo y la creación de la listas para la posterior construcción del dataset final.

Asimismo, mediante el uso de una estructura iterativa se extraen los datos que conformarán el dataset tomando como referencia atributos identificativos como la clase a las que estas pertenecen y se almacenan en las listas previamente creadas. Este procedimiento se realizará automáticamente con las cuatro primeras páginas que devuelve la búsqueda, dado que se pretende extraer una muestra de los hoteles disponibles. Finalmente, las listas formarán un dataset que quedará almacenado en el archivo Hotels Dataset.csv.

Una limitación que ofrece la plataforma, es que como técnica para prevenir el web scraping, la empresa cambia cada cierto tiempo los nombres de los elementos HTML vinculados a los datos. Consecuentemente, existe la posibilidad de que el código no funcione tras producirse dicho cambio. Por lo tanto, esto obliga al usuario a manipular el código introduciendo el nuevo nombre de las clases, los selectores o los Xpath. Por otro lado, otra limitación a tener en cuenta al enfocar este proyecto de scrapin, es que los links de acceso a la búsqueda concreta varían al menos una vez al mes, lo que genera un *bug* en el programa obligando a actualizar el link manualmente.

Por último, pueden consultarse las librerías utilizadas para el proyecto en el archivo requirements.txt contenido dentro del repositorio de Github.

10. Dataset

Se puede consultar tanto el dataset como su descripción en Zenodo mediante el siguiente enlace DOI : https://doi.org/10.5281/zenodo.7295560

11. Video

El video puede encontrarse publicado en Google Drive mediante el siguiente enlace: https://drive.google.com/file/d/17Dux8bg1pnLMfYS8wEq0f6gAQY75hNbQ/view?usp=sharing

12. Bibliografía

[1] INE (2022). Instituto Nacional de Estadística. *Estadística de movimientos turísticos en frontera. Frontur. Últimos datos*. Recogido de: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176996&menu=ultiDatos&idp=1254735576863

[2] Exceltur (2022). PIB Exceltur. https://www.exceltur.org/pib-turistico-espanol/.

[3] HOSTELTUR (2019). El turismo es el sector que más riqueza aporta a la economía española. https://www.hosteltur.com/130893_el-turismo-el-sector-que-mas-riqueza-aporta-a-la-economia-espanola.html

[4] eDreams (2022). Búsqueda de hoteles en Barcelona para el 30/11/2022. https://hotels.edreams.es/searchresults.es.html?aid=350432&label=edr-es-sbcitypairs-conf-pc-of&sid=49e883db440126547aaf71494c2245c0&sb=1&src=searchresults&src_elem=sb&error_url=https%3A%2F%2Fhotels.edreams.es%2Fsearchresults.es.html%3Faid%3D350432%26label%3Dedr-es-sbcitypairs-conf-pc-of%26sid%3D49e883db440126547aaf71494c2245c0%26tmpl%3Dsearchresults%26checkin_month%3D6%3Bcheckin_monthday%3D7%3Bcheckin_year%3D2023%3Bcheckout_month%3D6%3Bcheckout_monthday%3D9%3Bcheckout_year%3D2023%3Bcity%3D-372490%3Bclass_interval%3D1%3Bdest_id%3D-372490%3Bdest_type%3Dcity%3Bdtdisc%3D0%3Bfrom_sf%3D1%3Bgroup_adults%3D1%3Bgroup_children%3D0%3Binac%3D0%3Bindex_postcard%3D0%3Blabel_click%3Dundef%3Bno_rooms%3D1%3Boffset%3D0%3Bpostcard%3D0%3Broom1%3DA%3Bsb_price_type%3Dtotal%3Bshw_aparth%3D1%3Bslp_r_match%3D0%3Bsrc%3Dsearchresults%3Bsrc_elem%3Dsb%3Bsrpvid%3D666936fa8de70203%3Bss%3DBarcelona%3Bss_all%3D0%3Bssb%3Dempty%3Bsshis%3D0%3Bssne%3DBarcelona%3Bssne_untouched%3DBarcelona%26%26&ss=Barcelona&is_ski_area=0&ssne=Barcelona&ssne_untouched=Barcelona&city=-372490&checkin_year=2022&checkin_month=11&checkin_monthday=30&checkout_year=2022&checkout_month=12&checkout_monthday=1&group_adults=1&group_children=0&no_rooms=1&from_sf=1

[5] eDreams.Términos y condiciones eDreams https://www.edreams.es/images/shared/pdf/ES/flight_conditions.pdf

[6] Creative Commons. Attribution-ShareAlike 4.0 International
https://creativecommons.org/licenses/by-sa/4.0/legalcode

Contribuciones	**Firma**
Investigación previa	A.A O.G.
Redacción de las respuestas	A.A O.G.
Desarrollo del código	A.A O.G.
Participación en el video	A.A O.G.