Dataset: Habitaciones de hoteles en las 10 mejores ciudades del mundo

Truman Tapia Mora / Julio Zamora Guerrero

Contexto

El objetivo de la práctica es la creación de un dataset a partir de los datos contenidos en un sitio web. Para llevar esto a cabo hemos elegido Hoteles.com (https://es.hoteles.com/). Hoteles.com es uno de los principales proveedores mundiales de alojamiento hotelero y alquiler vacacional que ofrece servicios de reserva a través de su propia red de sitios web localizados, permitiendo acceder de manera centralizada a información sobre alojamientos, precios, servicios y disponibilidad.

Hotels.com L.P. es una de las compañías de viajes que conforman Expedia Group.

Título

Habitaciones de hoteles en las 10 mejores ciudades del mundo

Descripción del dataset

El conjunto de datos generado reúne atributos de habitaciones de hoteles en las 10 mejores ciudades del mundo de acuerdo con el ranking THE WORLD'S BEST CITIES 2021 (https://www.bestcities.org/rankings/worlds-best-cities/). Los datos cubren el intervalo de tiempo que va desde el 1/12/2021 hasta el 28/2/2022.

Representación gráfica

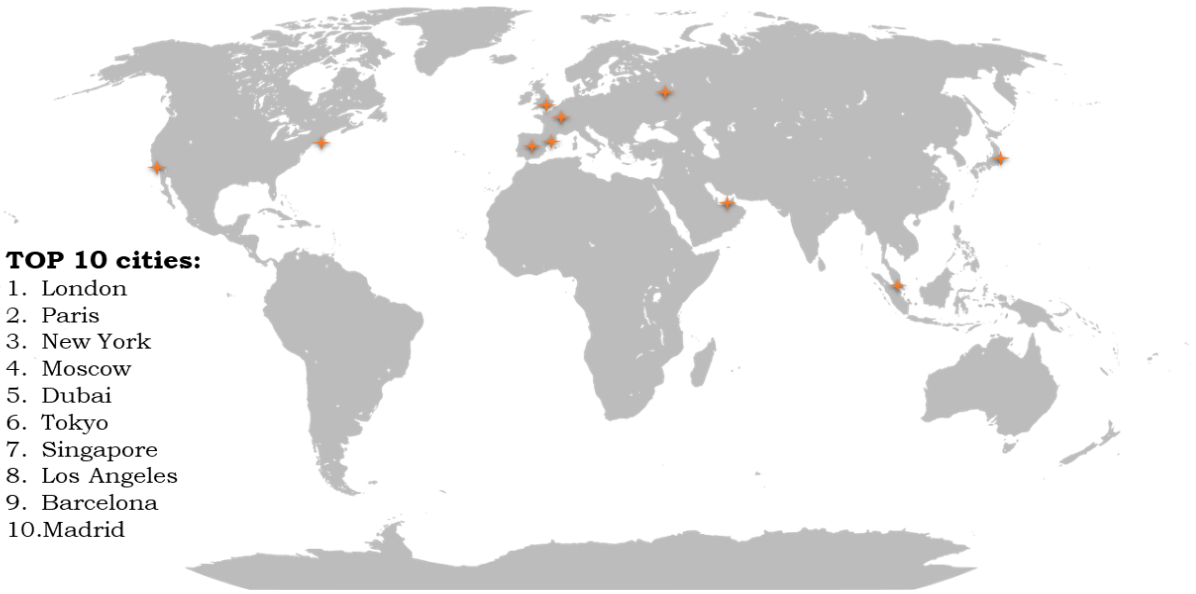

Ilustración 1. Top 10 ciudades según el ranking THE WORLD'S BEST CITIES 2021

Contenido

Cada fila del dataset corresponde a una habitación de hotel. Estos son sus atributos:

- **ciudad_busqueda** Ciudad del hotel
- **fecha_busqueda** Marca de tiempo de la búsqueda
- **fecha_registro** Fecha de entrada al hotel
- **nombre_establecimiento** Nombre del hotel
- **direccion_establecimiento** Dirección del hotel
- **servicios_establecimiento** Servicios disponibles del hotel
- **imagen_establecimiento** URL a la foto de portada del hotel
- **link_reserva** Link para reservar en hoteles.com
- **precio** Precio del hotel en USD por habitación y por noche

El rango del atributo **fecha_registro** va desde el 1/12/2021 hasta el 28/2/2022.

Agradecimientos

Los datos han sido recolectados desde Hoteles.com haciendo uso del lenguaje de programación Python, tecnología Wolfram y de técnicas de Web Scraping para extraer la información alojada en los HTML de los enlaces de búsqueda por localización y fecha.

Inspiración

El presente conjunto de datos puede servir como objeto de estudio para estudiantes y analistas de datos, ya que puede ser referencia para distintos proyectos en el ámbito de la ciencia de datos, como por ejemplo:

- Análisis exploratorio de los precios por localización y fecha de check in.
- Aplicación de técnicas de aprendizaje no supervisado (clustering de precios).
- Aplicación de técnicas de aprendizaje supervisado, para predicción de precios de habitaciones en distintas fechas.

También, el código provisto en el repositorio Github asociado a este dataset puede ser utilizado en la monitorización periódica de precios de hoteles. Por ejemplo, puede ser el caso de una aerolínea que frecuenta los mismos hoteles en sus principales destinos, y la información proveniente de la monitorización de precios puede resultar en un ahorro económico.

Licencia

La licencia escogida para la publicación de este conjunto de datos ha sido CC BY-NC-SA 4.0 License, cuyas cláusulas principales son:

- Atribución. Se debe dar crédito al nombre del creador del conjunto de datos generado, indicando los cambios que se han realizado. De este modo se reconoce el trabajo ajeno.
- No Comercial. No se puede usar el dataset para finalidades comerciales.
- Share Alike. Las contribuciones realizadas a posteriori sobre el trabajo publicado bajo esta licencia deberán distribuirse bajo la misma.

Una de las razones principales para elegir esta licencia es la cláusula No Comercial, para no interferir en el modelo de negocio de Hoteles.com.

Código

https://github.com/truman-t/Scrape-hoteles.com

Dataset

https://doi.org/10.5281/zenodo.5650235

Recursos

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

- Wolfram Research (2019), WebExecute, Wolfram Language function, https://reference.wolfram.com/language/ref/WebExecute.html.

Tabla de contribuciones

Contribuciones	Firma
Investigación previa	TTM; JZG
Redacción de las respuestas	TTM; JZG
Desarrollo del código	TTM; JZG