Tipología y Ciclo de Vida de los Datos: PRÁCTICA 1

Autores: Jorge Ramón Díaz Suarez y Víctor Fernández Moreno

Noviembre 2022

Índice

Memoria

En este archivo se responderán a las preguntas que son planteadas en el enunciado de la práctica.

Todos los archivos pertenecientes a esta práctica se enceuntran en el repositorio de GitHub **M2.851 - PRÁCTICA 1**.

Contexto

El caso que planteamos en esta práctica, es el de una compañía que se encarga de organizar eventos, es por ello, que si quiere tener un mayor alcance y captar más clientes, tendrá que facilitar el viaje a aquellos que no vivan en la misma ciudad en la que se organiza dicho evento, para ello, se facilitará a los clientes un listado de hoteles que se encuentren cerca del recinto en el que se organiza el propio evento. Para llevar a cabo esta idea, se decide extraer la información, mediante el uso de *Web Scraping* en Python, de una famosa página web de reseñas de hoteles como es **TripAdvisor**.

Dataset

Título

Hemos decidido extraer dos .csv que son de interés para el objetivo que queremos llevar a cabo. El primero, **df_hoteles.csv** en el se encuentran los datos almacenados en forma de dataframe, cada fila hace referencia a un determinado hotel, y en las columnas encontramos los distintos atributos de los hoteles. El segundo .csv, tiene el nombre de **df_comentarios.csv**, este es otro dataframe, en el que en cada fila hay información

relativa a un comentario publicado por un cliente, y en las columnas encontramos carácteristicas de dicho comentario.

Descripción del dataset

Las distintas variables que encontramos dentro del dataset **df_hoteles.csv** son:

- **nombre**: Hace referencia al nombre del hotel.
- **ciudad**: Ciudad en la que se encuentra el hotel.
- **direccion**: Hace referencia a la dirección del hotel.
- **descripcion**: Breve descripción a cerca del hotel.
- **precio**: Lista con los precios a los cuales se puede reservar una habitación.
- **amenities**: Cadena de texto con todas las facilidades que proporciona el hotel.
- **puntuacion**: Valoración del hotel según los clientes.

Aquí cabe destacar que la variable *amenities*, puede parecer a primera vista poco útil, aunque tras un tratamiento de procesado de texto, puede ser una de las variables más relevantes del dataset, ya que a partir de ella se podrán construir otras variables dicotómicas en las que queda reflejado si el hotel dispone de un servicio o no.

Por otro lado, la información que se obtiene es la que se encuentra actualmente disponible en la web de TripAdvisor, sin hacer restricciones por fecha, por lo que se proporciona un listado global de hoteles. Una manera de hacer más compleja esta tarea sería proporcionar únicamente los hoteles disponibles en un período de tiempo, pero esto supone la dificultad de que tenemos que scrapear la web periódicamente para actualizar la información, añadiendo los hoteles con habitaciones diponibles y descartando los hoteles que se hayan quedado sin habitaciones.

Mientras que las distitnas variables que encontramos en el dataset **df_comentarios.csv** son:

- **titulo**: Hace referencia al título del comentario.
- **contenido**: Se trata del comentario como tal.
- **autor**: Persona que ha escrito el comentario.
- **hotel**: Hotel al que hace referencia el comentario.
- **calificacion**: Puntuación que da el cliente al hotel.

El motivo por el cual se ha decido extraer este dataset, es para ampliar la información que se puede extraer de cada hotel, ahora desde una perspectiva de cliente. En el tratamiento de los datos, será interesante realizar un procesado de texto que nos ayude a encontrar en los comentarios ciertas palabras que nos proporcionen cierta información del hotel, un par de ejemplos de palabras que se pueden buscar en estos comentarios serían: "habitación limpia" o "buena comida".

Representación Gráfica

Se añade como anexo al final del documento.

Propietario

El propietario del dominio de TripAdvisor es la compañia: **TripAdvisor LLC**.

Esta información se ha conseguido gracias al paquete de python llamado *whois*, este proporciona información de itnerés como el propietario del domino, pero también ofrece otros datos que también pueden ser relevantes como a quien le pertenece el dominio o cuando se ha actualizado la página.

Inspiración

Al comienzo de este proyecto los integrantes de este grupo estuvimos discutiendo sobre que datos eran los más adecuados para realizar esta tarea, se plantearon distintas opciones para realizar web scraping pero todas ellas se descartaron finalmente por tratarse de casos muy concretos y de poco interés para el lector. Por lo que se decidio seguir un camino más cotidiano, por ello se decicidó extraer la información de los hoteles.

Como se expuso en el contexto, al decidir usar esta temática en concreto, se plantea resolver el problema que pueda tener una empresa que se encargue de organizar eventos a cerca de como aumentar su público objetivo, para ello tendrán que hacer todo lo posible para facilitar a los asistentes el viaje y la estancia, en nuestro caso nos hemos centrado en la estancia. Es por ello que busca responder a preguntas del estilo, ¿Qué hoteles hay en la zona?¿Qué valoración tiene ese hotel según los clientes?¿Qué facilidades ofrece dicho hotel? ó ¿Por cuánto se puede reservar una habitación?, motivados por estas preguntas decidimos seleccionar las variables anteriores, las cuales darán respuestas a las preguntas planteadas.

Licencia

Tras haber consultado las condiciones y términos de uso del sitio web, llegamos a la conclusión que la licencia que mejor se adapta a nuestras necesidades es: **Released Under CC BY-NC-SA 4.0 License**.

En las condiciones de uso de la web se explica no esta permitido el uso de robots, scrapeadores o cualquier otra herramienta similar, si el objetivo de extraer estos datos es comercial, es por eso mismo que para nuestros objetivos no estamos incumpliendo las condiciones del sitio web y la licencia seleccionada es la más adeucada.

Código

En el repositorio de la práctica, además de los scripts, se encuentra el archivo **requirements.txt** en el cual se encuentra información a cerca de la versión de python usada y las distintas librerias de las que se hacen uso.

El código que se ha realizado en esta práctica se trata de un código bastante estándar de web scraping en el que se ha utilizado el paquete *Scrapy*.

La dificultad, no se encuentra en el código como tal, ya que no estamos ante un código excesivamente complejo, los principales problemas se encuentran a la hora de extraer información de la página, a la hora de inspeccionar el código fuente de esta y seleccionar los elementos que son de interés.

Zenodo

El DOI de Zenodo es: **10.5281/zenodo.7324105**

Vídeo

El enlace al vídeo explicativo de la práctica se encuentra en el enlace sigueinte: **https://drive.google.com/file/d/1Fl-b_Gv-hc8-bBWStA-lyRioJCcVWAIs/view?usp=share_link**

Contribuciones

Contribuciones	Firma
Investigación previa	Jorge Ramón Díaz Suarez, Víctor Fernández Moreno
Redacción de las respuestas	Jorge Ramón Díaz Suarez, Víctor Fernández Moreno
Desarrollo del código	Jorge Ramón Díaz Suarez, Víctor Fernández Moreno
Participación en el vídeo	Jorge Ramón Díaz Suarez, Víctor Fernández Moreno

Anexo

Figure 1: Representación gráfica del proyecto