PRA1: Web Scraping TripAdvisor con Python

Alumno: Brais Rodríguez Martínez

Nombre de usuario: brodmar

Contexto

En el papel de una cadena de hoteles a nivel de España, se quiere llevar a cabo una tarea de extracción de información acerca de hoteles según la ciudad de interés con el objetivo de observar a la competencia y en qué hace más hincapié a la hora de dejar una opinión. Además, usando los comentarios, se puede realizar un procesado de estos y entregárselos a un modelo que permita detectar si es un buen o mal comentario en esta y otras páginas web. Alternativamente, se podría llegar a detectar si es o no un comentario verídico.

Por ser una fuente de información fiable y abundante, la información recogida proviene de TripAdvisor donde, dependiendo de la zona de análisis, se recogen los hoteles más populares de la zona con sus datos destacados y se accede al enlace de cada uno de ellos, extrayendo los comentarios que deja la gente.

Título

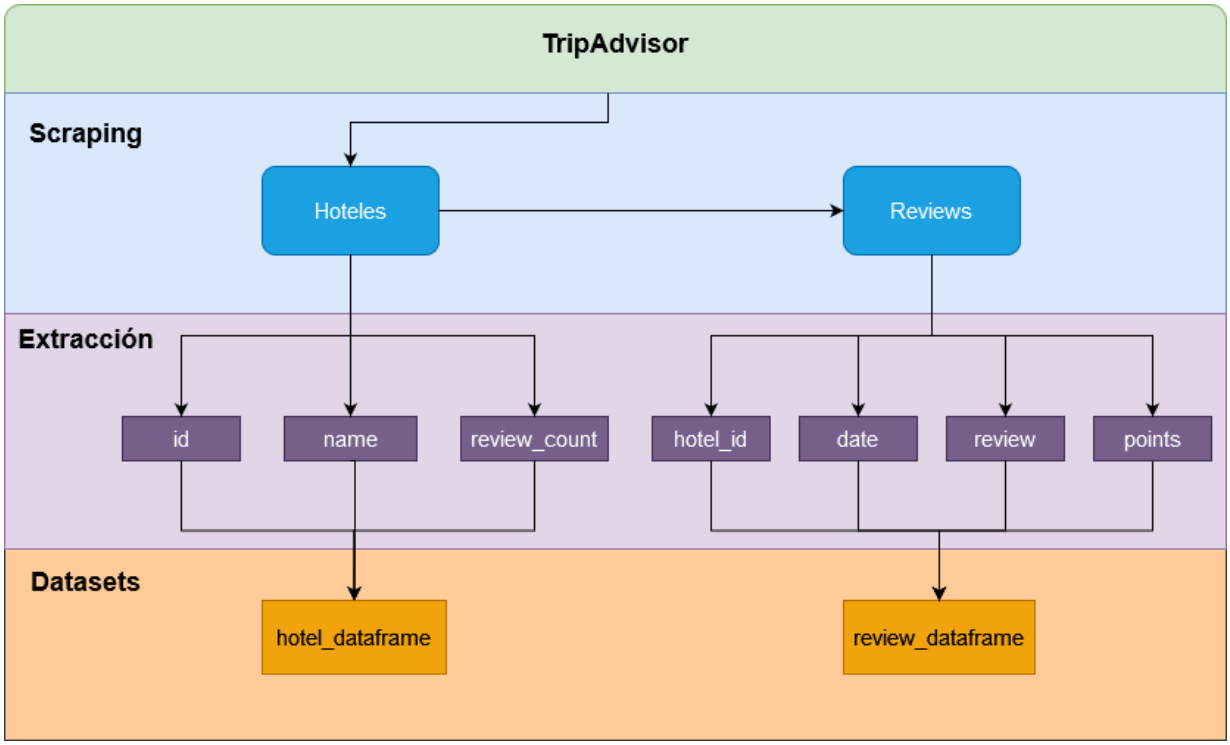
Datos de hoteles y sus opiniones.

Descripción del dataset

El dataset contiene datos de hoteles presentes en la web de TripAdvisor y sus opiniones. Por una parte, se han guardado los hoteles más destacados de una ciudad en concreto, recogiendo su nombre, un id generado para cada uno y el número de opiniones totales. Por otra parte, todas las reviews del establecimiento, incluyendo la fecha en la que se publicó, el id generado del hotel y la puntuación que la persona puso al mismo.

Representación gráfica

A continuación, se muestra un diagrama de flujo del proyecto:



Se destacan principalmente 3 fases:

- **Scraping:** el propio proceso de web scraping de la web. Se accede a la zona elegida y se extrae el HTML de los hoteles y, accediendo a la URL de cada uno de ellos, el de las reviews.
- **Extracción:** de cada hotel y de cada review del mismo, se extraen los atributos indicados, que son los elegidos provisionalmente para el dataset final.
- **Datasets:** con la información extraída, se montan dos CSV, uno conteniendo todos los hoteles y otro todas las reviews, que son guardados en el directorio del proyecto.

Contenido

El dataset final consta de los siguientes atributos:

- id: identificador numérico incremental del hotel.
- name: nombre del hotel que consta en la web.
- review_count: número total de opiniones totales del hotel.
- date: mes y año en el que se publicó la opinión
- review: opinión literal que el usuario ha dejado como comentario después de su estancia en el hotel.
- points: nota numérica sobre 5 que el usuario le ha puesto al establecimiento.

La recogida de datos tiene como objetivo una recopilación histórica, desde el primer comentario hasta el último, ya que se le quiere dar más prioridad al propio texto escrito por la persona.

Relativo al proceso de recogida de datos, primero se selecciona la ciudad de la que se quiere extraer los datos, inicialmente Ourense. Se recorre la lista de los 30 hoteles más destacados

por la web, recogiendo información básica que aparece en cada una de las pequeñas tarjetas que la web usa para presentar los datos y, para cada uno de ellos, se accede a su URL y se hace un recorrido anidado a través de todas las opiniones, de manera similar a cómo se ha hecho con los hoteles. En cada apartado de opinión, se recoge la información descrita. Una vez se ha terminado el proceso, toda la información se guarda en 2 CSV, uno para hoteles y otro para las opiniones, con el objetivo de no repetir la información del hotel.

Agradecimientos

Principalmente, los agradecimientos van dirigidos a la web TripAdvisor, ya que todos los datos provienen de ahí. Sobre todo, lo mejor ha sido la ausencia de métodos de contención agresivos por parte de la web para extraer información, limitándose a establecer una tasa de 50 peticiones por día y 5 por segundo, que se ha respetado en todo momento.

En relación a la propia extracción de los datos, se ha revisado el archivo robots.txt y se han revisado todas las URLs a las que la página tiene prohibido el acceso, siendo todas rutas relacionadas con ámbitos internos y de pago de la página que podrían no respetar los criterios de privacidad y seguridad del sistema. Además, se ha revisado la lista de bots bloqueados por la web, aunque sin poner tanta atención debido a que no se está usando ningún bot. Por otra parte, para respetar la ratio de peticiones por día y segundo, se ha establecido un delay aleatorio de entre 1 y 5 segundos entre cada petición.

Para la realización de este proyecto, se ha realizado una investigación de los análisis ya existentes relacionados con esta página. Se destaca que **no se ha realizado ningún tipo de plagio acerca del código** usado en los siguientes enlaces, lo que salta a la vista si se comparan con el código desarrollado. **Sólo ha sido usado para ver el actual contexto del web scraping de TripAdvisor en los últimos años**:

- *Sentiment-Analysis-on-TripAdvisor-reviews* de gabrieletiboni: ha sido la inspiración para la realización de un proyecto de este tipo, una vez la idea se puso sobre la mesa, se ha realizado una investigación y se ha encontrado un análisis como este. Se trata de un script y documentación completa para realizar análisis de sentimiento a partir de reviews de TripAdvisor y análisis descriptivo de las palabras más usadas. El repositorio es el siguiente https://github.com/gabrieletiboni/Sentiment-Analysis-on-TripAdvisor-reviews/tree/master/CODE.
- *Scraping TripAdvisor with Python 2020* de giuseppegambino: una vez se había planteado la idea, se ha revisado si había alguna consideración a tener en cuenta para realizar el web scraping de TripAdvisor, como algún tipo de control del flujo de reviews por segundo o las posibilidades del mismo scraping, como realizarlo teniendo como referencia restaurantes, cosas que hacer, hoteles, etc. El código y la documentación muestran un proyecto que realiza web scraping de restaurantes y cosas que hacer, repositado en: https://github.com/giuseppegambino/Scraping-TripAdvisor-with-Python-2020.

Inspiración

Con la información recogida en el dataset final y tomando de inspiración el análisis ejecutado del apartado 6, se pretende realizar una serie de análisis que permitan a una cadena de hoteles sacar conclusiones acerca de qué es lo que le preocupa más a la gente, qué le molesta o qué más gusta o simplemente acerca de qué se habla más.

Esto se logra con un análisis descriptivo de las opiniones, recogiendo las palabras más usadas que signifiquen algo, descartando las llamadas *stopwords*, palabras que no aportan nada al análisis como demostrativos, posesivos, pronombres, etc. Acompañándolo de la puntuación que la persona ha puesto para saber un poco más acerca del contexto del uso de esas palabras.

Por otra parte, a modo de análisis mucho más avanzado, se podría entrenar a un algoritmo de clasificación que permita identificar el sentimiento de la opinión para saber si la experiencia ha sido buena o mala.

Todo esto se puede aplicar para el beneficio del propio hotel, de manera que invierta más atención y dinero en los aspectos más comunes, pero también puede ser usado para analizar la competencia y, en lugar de hacer un análisis general, hacerlo más específico de lo que falla en los hoteles de la competencia.

Licencia

Debido a que se trata de información guardada en una página web pública accesible a todo el mundo, el dataset final está sometido a una licencia “Released Under CC0: Public Domain License”.

Código

Todo el código usado para generar el dataset final se encuentra en el repositorio github https://github.com/rodriguezbrais/pec1. En él, se puede encontrar el archivo python hotel_scraping, así como un archivo readme que explica más en profundidad el código desarrollado.

Dataset

El dataset final se encuentra subido a la plataforma Zenodo y accesible a través del siguiente enlace https://zenodo.org/record/5614241.

Tabla de contribuciones

Contribuciones	Firma
Investigación previa	B.R.M. (Brais Rodríguez Martínez)
Redacción de las respuestas	B.R.M. (Brais Rodríguez Martínez)
Desarrollo del código	B.R.M. (Brais Rodríguez Martínez)