Práctica de Open data y visualización dinámica

Documento Explicativo

Juan Pablo Campuzano Medina

04/02/2024

A continuación, se plantea un pequeño documento explicativo del scrapper y visualización de los datos obtenidos de la sección de restaurantes de TripAdvisor España, donde extraemos los restaurantes mejor puntuados de las 20 ciudades españolas más pobladas

1. Extracción:

Fichero scraper_TripAdvisorCampuzanoMedina.py
Herramientas usadas:

- **Selenium:** Usamos esta librería debido a que es necesario interactuar con el sitio web (botón de cookies, introducción de texto) y debido a que las URL no siguen una estructura automatizable
- **BeatifulSoup:** Usamos esta librería ya que el HTML parser es una herramienta muy útil para ciclar entre elementos de una web

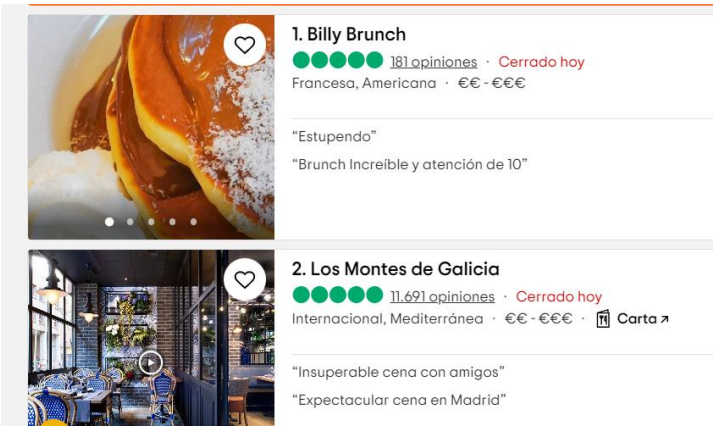
Flujo de funcionamiento del script (para mayores detalles en los comandos usados hacer referencia a los comentarios del script)**:**

- Importamos las librerías necesarias para la extracción y la escritura de datos
- Hacemos un loop que nos permite buscar varias ciudades
- Para cada ciudad hacemos lo siguiente:
  - Abrimos https://www.tripadvisor.es/Restaurants

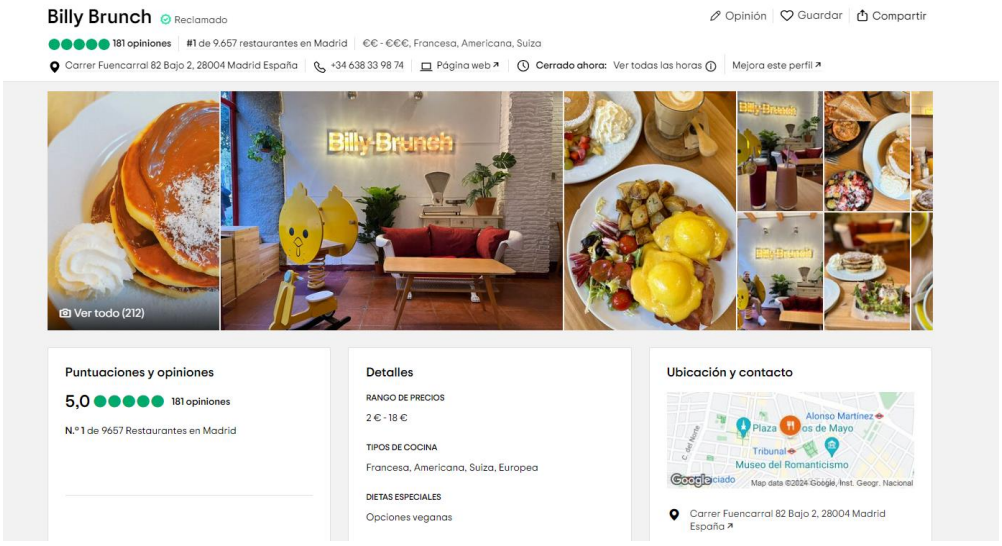
  - Aceptamos el botón de cookies

- Introducimos letra por letra el nombre de la ciudad y seleccionamos la primera opción sugerida, esto es importante ya que es el modo más correcto de delimitar geográficamente la búsqueda

Buscamos la sección del HTML donde se encuentran los restaurantes, ignorando ofertas y sugerencias, usamos el filtro por defecto que es la puntuación más alta.



- Para Cada restaurante de los mostrados en la página, usando beatifulsoup extraemos el url y abrimos en una nueva pestaña la página dedicada al restaurante



- Extraemos la siguiente información:
    - - Name: Nombre del Restaurante
    - - Reviews_N: Numero de reseñas
    - - Rank: Rango en su ciudad
    - - Price: Precio
    - - Categories: Tipos de cocina ofertada
    - - Adress: Dirección del restaurante
    - - Score: Puntuación del restaurante
    - - Latitude: Latitud de la ubicación
    - - Longitude: longitud de la ubicación
    - - City: Ciudad donde está ubicación
- Cerramos la pestaña y repetimos la extracción por cada uno de los restaurantes de la pagina
- En caso de cargar más de una página se busca el botón de página siguiente hasta alcanzar el número de paginas
- Repetimos estos dos ciclos (restaurantes dentro de página y número de páginas) por cada una de las ciudades.
- Cerramos el driver

Ejecución del Script: para ejecutar el script verificar de tener instaladas las librerías mencionadas anteriormente al igual que el navegador google Chrome, el scrapper no funciona en modo navegador 'Headless'

Pip install selenium

Pip install BeautifulSoup

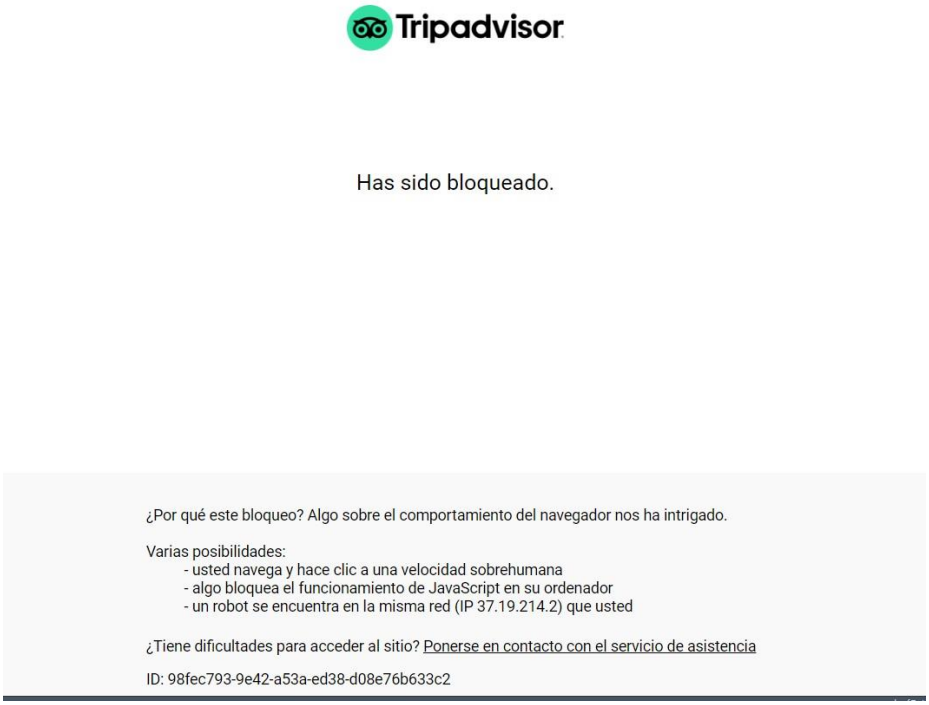
Python run scraper_TripAdvisorCampuzanoMedina.py

Para modificar las ciudades cambiar 'spain_cities' dentro del fichero

Para modificar el número de páginas a scrapear modificar 'num_pages_to_download'

Limitaciones del scrapper:

Después de varios días desarrollando el scrapper, el sitio web detecto nuestro bot. Para hacerlo funcionar tuve que cambiar de ordenador y usar una VPN.



Otra limitación es la velocidad de Selenium que es poco lento para extraer toda la información ya que el tiempo de ejecución cargando una página por ciudad fue cerca de 30 min.

De igual forma el uso de memoria RAM por parte se Selenium en combinación con Chrome es elevado, por lo que seguramente hay espacio para optimización del script

2. Procesamiento:

El Procesamiento de los datos se hizo en 3 partes diferentes:

Dentro del Scrapper: Al extraer la información se prepararon los datos con el tipo correcto además se obtuvieron la latitud y longitud extrayéndolas del string presente en el mapa embebido dentro del sitio web usando expresiones regulares.

Dentro del archivo de visualización (Viz_dinamic_campuzano_Tripadvisor.py):

- Se recodifico la columna precio cambiando los símbolos de euro por la escala: bajo, medio, alto

- Se removieron los duplicados, usando la columna nombre, la presencia de duplicados se debe a la estrategia publicitaria de tripadvisor, de meter el mismo restaurante en diversas posiciones en una misma página.
- Limpieza de la columna del número de reseñas que estaban dentro de u la frase ‘1115 opiniones’
- Limpieza de la columna score, remplazando el separador decimal y haciendo la el cast a float

Dentro del archivo con las funciones para realizar los gráficos (plots_dinamic.py):

- Se extrajeron a cada una de las categorías de los restaurantes (que pueden ser varias por cada uno, usando la función .explode(), para poder filtrar por categoría de cocina de forma correcta y realizar la gráfica correspondiente
- Se agrupo por categoría y ciudad y se hizo el conteo para poder realizar el histograma agrupado por ciudad
- Se agrupo por precio y ciudad y se hizo el conteo para poder realizar el histograma agrupado por ciudad

3. Justificación de Visualizaciones.

Página de Datos:
Es importante presentar el datagrama al usuario de modo ordenado, usando la herramienta interactiva de streamlit que permite ordenar las columnas, además con la introducción de filtros variados, el usuario podrá introducir sus preferencias de cocina y ubicación para obtener los diferentes restaurantes y ordenarlos a conveniencia.

Página de visualización:
Teniendo en cuenta que se extrajeron los 30 mejores restaurantes de cada ciudad el objetivo es poder analizar los tipos de cocinas que son más apreciados por el público de las diferentes ciudades.

Las categorías de los restaurantes se pueden explorar con el histograma no agrupado con el fin de comparar la popularidad de los diferentes tipos de cocina o se puede explorar agrupando por ciudades para comparar la cantidad de ofertas de diferentes cocinas por ciudad

De igual forma, otro factor determinante a la hora de elegir dónde comer, es el precio, por lo que hace un análisis similar al anterior, para comprar cuan costoso es comer en los mejores restaurantes de cada ciudad.

Para más detalles e insights hacer referencia a la visualización.

Página del Mapa:
El objetivo de esta página es simular un buscador de restaurantes, donde se puede filtrar por diferentes parámetros y elegir el restaurante que más se ajuste a las necesidades del usuario

Para Ejecutar la visualización es necesario estar en un directorio que contenga los siguientes ficheros:

- **plots_dinamic.py**
- **extracted_datafinal.jsonl**
- **Viz_dinamic_campuzano_Tripadvisor.py**
- **Tripadvisor.png**

Navegar con la terminar al directorio y ejecutar el comando

streamlit run Viz_dinamic_campuzano_Tripadvisor.py