Escuela de ingeniería informática, Universidad de Las Palmas de Gran Canaria

Proyecto Scraper

Desarrollo de Aplicaciones para Ciencia de Datos

Javier Franco González

Grado en Ciencia e ingeniería de datos, 2º curso.

Resumen

En este proyecto se ha realizado web Scraping de la página web de Booking.com, para obtener información sobre hoteles. El programa es capaz de obtener información de la ubicación del hotel, así como sus servicios, reseñas y valoraciones. También se ha implementado un método opcional para obtener los nombres de los hoteles en una ciudad determinada.

La mayor parte del código de esta práctica la encontramos en la clase Booking Scraper. Dentro de esta clase se encuentran los métodos para obtener separadamente la información comentada previamente. Estos métodos usan como parámetros el nombre del hotel y el código del país donde se encuentra, por ejemplo “es” si se trata de España, y devuelven los datos en formato json, o jsonarray si son varios. Estos métodos acceden a la url de hotel gracias a jsoup y crean un objeto de la clase Document. Dentro del documento se buscan los datos que queremos obtener para posteriormente extraerlos.

Para extraer los datos se busca un elemento que contiene la información requerida, no obstante no fue posible hacer esto en todos los casos. En varias de las funciones se tuvo que obtener manualmente (inspeccionando en la página web de Booking) la clase a la que pertenecían los datos que buscaba, y una vez obtenido el nombre de la clase elegir los elementos del documento que perteneciesen a dicha clase.

En la función de obtener las valoraciones, al obtenerlas varias aparecían repetidas, por lo que fue necesario poder crear una clase Valoración para poder compararlas entre sí y no añadir una nueva valoración si esta ya se había extraído previamente.

La función que obtiene el nombre de los hoteles de una ciudad funciona de la misma manera que las anteriores, pero conectándose a una url distinta y devolviendo un Array de Strings que corresponden a los nombres de los hoteles. En este caso fue más difícil, y la clase donde se encontraba el nombre de cada hotel tuvo que transformarse a String para poder obtener una substring que sería el nombre del hotel.

Por último encontramos la clase Main, desde donde se realizan las peticiones. En ella podemos encontrar un ejemplo donde obtiene los nombres de los hoteles de Madrid y obtiene las valoraciones de cada uno de ellos.

Índice:

Recursos utilizados

El entorno de programación que se ha utilizado es InteliJ. El SDK del proyecto es Oracle JDK versión 11. También me he documentado mediante numerosas páginas web que pueden encontrarse en la bibliografía de esta memoria.

Diseño

El diseño se ha centrado en presentar un código con nombres de variables que te ayudan a comprenderlo. No han sido necesarias muchas clases ya que el proyecto tenía una única funcionalidad. Se ha intentado que cada método mantenga la misma estructura ya todos funcionan con el mismo objetivo, de modo que sea más fácil de realizarlos y entenderlos.

Conclusiones

He encontrado muchas dificultades realizando esta práctica. La obtención de datos sin ayuda de una API me ha resultado difícil ya que no conocía el nombre de las variables que quiero obtener, y tampoco estaba seguro de a que url debía conectarme. Al final he aprendido a analizar inspeccionando el código de la web de Booking para poder obtener los datos, a pesar de que no creo que sea la manera más eficiente. Puede ser una habilidad que me resulte útil en el fututo, pero si realizase nuevamente el trabajo me gustaría encontrar un método más versátil de hacer Scraping.

Líneas futuras

Este programa puede mejorarse bastante y ser muy útil para toda persona que desee quedarse en un hotel. El programa podría obtener datos de muchas más páginas de hoteles además de Booking, y realizar comparaciones entre ellos. Se podrían obtener datos adicionales como la disponibilidad de cada hotel. Creando una buena interfaz podría resultar muy atractivo para los clientes usar este programa para de este modo conocer los hoteles de su destino turístico y realizar una buena elección.

Bibliografía

https://mvnrepository.com/

https://stackoverflow.com/

https://chat.openai.com/chat

https://www.booking.com/index.es.html?aid=304142&label=gen173nr-1FCAEoggI46AdIM1gEaEaIAQGYAQq4ARfIAQzYAQHoAQH4AQ2IAgGoAgO4Avfo2p0GwAIB0gIkYzZlZTFjYzMtMWFjMS00ZDYwLWFiOTUtM2M0Mzk4Yjg5N2Mz2AIG4AIB&sid=d5bd35482eefb3fc4e0af63c33031cf3&click_from_logo=1

https://www.javatpoint.com/jsoup-tutorial

Javier Franco González, 10/1/2023.