

Booking Scraper



Desarrollo de Aplicaciones para Ciencia de Datos.
Segundo año del Grado en Ciencia e Ingeniería de Datos.

Proyecto realizado por Gerardo León Quintana.

Fecha de realización de la memoria 05/01/2023.

Resumen

El proyecto consiste en la realización de un web scraper para obtener los datos de una página web basada en la reserva de hoteles, en este caso Booking.com.

En dicho proyecto se ha implementado una interfaz dedicada al scrapping de páginas web basadas en la reserva de hoteles, llamada HotelScraper, dicha interfaz posee cuatro métodos.

Debido a que se ha elegido para este proyecto a Booking.com, se ha creado una clase llamada BookingScraper la cual implementa los métodos de HotelScraper adaptados para el scraping de dicha página. Cada uno de esos métodos corresponde con los requisitos que debe poseer el proyecto a entregar, siendo el primero el método getHotelName, el cuál al pasarle el hotel elegido, nos devolverá el nombre del hotel, su ubicación y sus coordenadas. De seguido, tenemos el método getHotelServices, el cuál al pasarle el hotel elegido, nos devolverá cada uno de los servicios que ofrece el hotel introducido a la función. Posteriormente, tenemos el método getHotelComments, el cuál al pasarle el hotel elegido, nos devolverá la primera página de comentarios del hotel introducido. Finalmente, tenemos el método getHotelRatings, el cuál al pasarle el hotel elegido, nos devolverá las valoraciones del hotel introducido.

Además, este proyecto cuenta con una clase WebService el cual se encarga de inicializar una API con la ayuda de la librería Spark, con la cual nuestros clientes podrán visualizar el contenido obtenido a través del scraper. Para ello se ha creado el método start, el cual posee cuatro peticiones GET, cada una de ellas corresponde con uno de los métodos de HotelScraper, y para ello le hemos inyectado a WebService la interfaz HotelScraper, para así poder utilizar cualquier tipo de clase que implemente los métodos de HotelScraper.

Finalmente se ha creado una clase Controller la cual se encarga de llevar a cabo todos los procesos del programa, todo ello para obtener una ejecución lo más organizada posible.

Índice

- Página 5:
 - Recursos utilizados.
 - Diseño.
 - Diagrama de clase.
 - Líneas futuras.
- Página 6:
 - Conclusiones.
 - Bibliografía

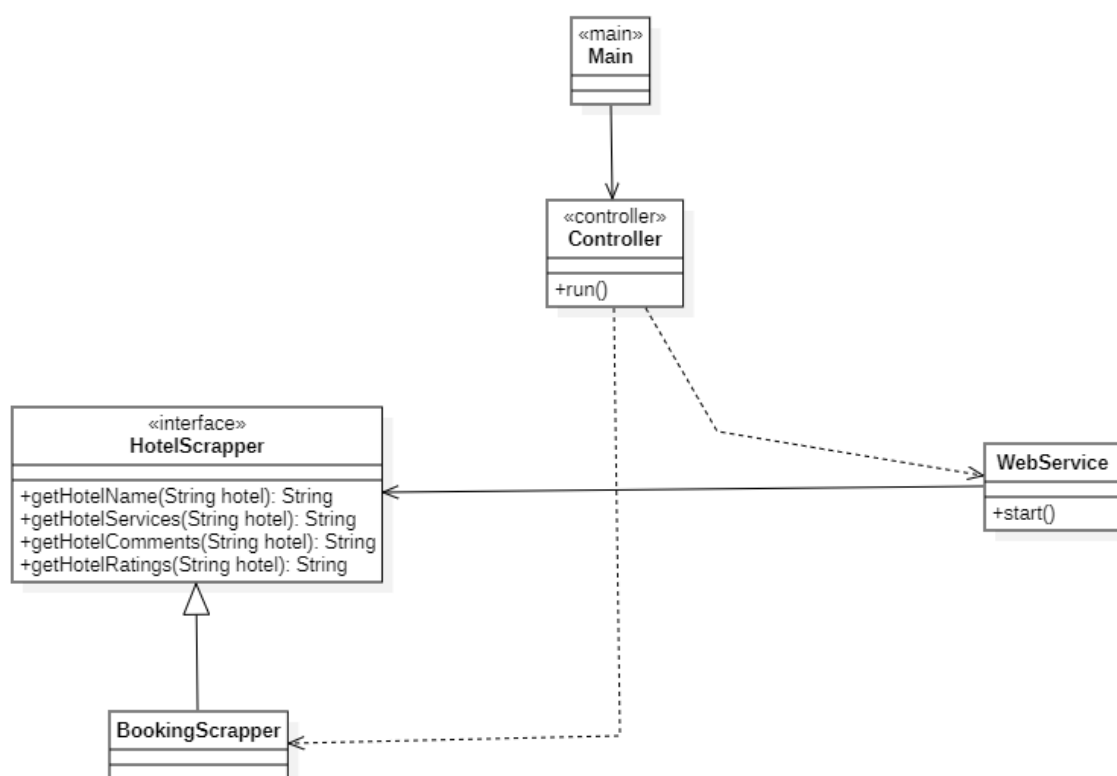
Recursos utilizados

La realización íntegra del proyecto se ha realizado en IntelliJ.

Diseño

El diseño de este proyecto se considera como un MVC, donde el modelo correspondería con la interfaz HotelScrapper y a las clases que la implementan, la vista correspondería con la clase WebService y el control con la clase Controller.

Diagrama de clase



Líneas futuras

Se podría crear una factoría de scrappers con los cuales realizar el scraping con distintas páginas de reserva de hoteles. Para ello implementaríamos un nuevo parámetro en los paths de los métodos GET del WebService, con el cual el cliente nos indique la página web con la cual quisiera realizar el scraping.

Conclusiones

La realización de este proyecto ha supuesto un reto para mí, esto se debe a que para realizar scraping hace falta una base de conocimientos de HTML, además aprender a usar de forma correcta la librería JSOUP, la cual nos permite realizar todo el proceso del scraping en java.

Bibliografía

- <https://sparkjava.com/>
- <https://jsoup.org/>
- <https://www.geeksforgeeks.org/how-to-setup-jackson-in-java-application/>