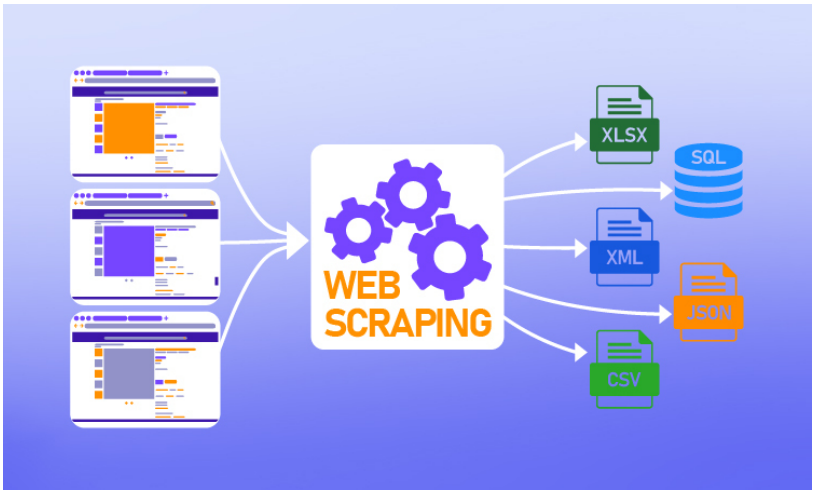
PROYECTO SCRAPPER A API REST

Desarrollo de Aplicaciones
para Ciencia de Datos

2º Ciencia e Ingeniería de Datos

Escuela de Ingeniería Informática - Universidad de Las Palmas de Gran Canaria (ULPGC)



Resumen

El proyecto consiste en extraer información de hoteles de la página web de booking, y mostrar el resultado del web scraping mediante una api rest que creamos gracias a Spark. El web scraping se realiza en la clase Scrapper, donde se extraen datos de 4 principales elementos: ubicación del hotel, servicios, calificación y reviews (este último realizado de la página booking reviews, del cual solo obtenemos los 25 comentarios más destacados del idioma español, debido a que si scrapearamos todos los comentarios de todas las páginas de comentarios de todos los idiomas, el programa tardaría en ejecutarse muchísimo tiempo). Estos han sido almacenados en listas de objetos, para posteriormente en la clase ApiRestController ser filtrados por nombre del hotel al que pertenecen para que funcionen correctamente los formatos de respuesta. Cabe destacar que a parte de los 4 requisitos que se nos pedía, en este proyecto se ha añadido un método get con la ruta “/hotels”, en la que la respuesta es la lista de hoteles de la que se dispone.

Recursos utilizados

-Intellij
-Git
-GitHub
-Google
-Spark
-Google Docs
-Talend API Tester
-Jsoup
-Maven
-Maven central

Diseño

Model view controller (Sin view porque no se crea interfaz)

Líneas futuras

La idea de negocio en este caso es clara, se puede vender el programa a cualquier persona que quiera consultar distintas opciones de hoteles a los que visitar.

Conclusiones

Lecciones aprendidas: Gracias a este proyecto ahora me ha quedado muy claro cómo usar Spark y cómo funcionan las Api Rest. El web scraping ya sabía hacerlo porque para la asignatura “Fundamentos de Marketing y Comportamiento del Consumidor” ya me pidieron hacer uno, pero aproveché y ya lo hice con jsoup. El código de este trabajo tiene algunas partes sacadas del código utilizado de aquella vez, lo cual ahorró tiempo (el código del web scraping de las reviews está en gran parte basado de ese trabajo) permitiendo estar mucho más relajado para el resto de desafíos que ofrecía este proyecto. Otra lección aprendida es que gracias a la corrección del anterior trabajo ahora se que tengo que inyectar la lista con los nombres de los hoteles por si en algún momento se tiene que modificar y es lo que he intentado plasmar.

Si empezara de nuevo: Debería hacer las pruebas con 2 hoteles, en vez de con los 10 ya que me ahorraría mucho tiempo de compilación.

Bibliografía

https://sparkjava.com/
https://www.booking.com/
https://jsoup.org/

Joaquín Ibáñez Penalva
09/01/23