8-11-2024

Captura y Preparación de Datos.

PEC2

Juan Luis Acebal Rico
GRADO DE CIENCIA DE DATOS APLICADA

Indice

Parte teórica

Q1.

¿Qué es el archivo "robots.txt" y por qué es importante analizar su contenido? ¿Qué riesgo estamos tomando si no lo hacemos?

Es un archivo de texto plano que está en el directorio raíz de un sitio web, y que indica a los bots, rastreadores de contenido (por ejemplo, motores de búsqueda, scrapers, etc), lo que se puede o no hacer. Es importante saber en un entorno de producción (en todos, pero especialmente en un entorno de producción), qué podemos hacer o no.

Un ejemplo propio, es que personalmente he intentado hacer web scraping en amazon y linkedin con motivos didácticos, y si bien de base está prohibido y por algo así no me puedo meter en problemas, a priori, aun no siendo legal hacerlo, en entornos profesionales, es también no legal, alegal, con serias dudas de su legalidad, o ilegal, y puede ser una fuente de problemas. Por tanto es muy importante ver el archivo robots.txt antes de realizar scraping.

Q2.

¿Cuándo realizamos una petición HTML, qué tipo de errores puede devolver el servidor y en qué grupos pueden ser clasificados?

Puede ser clasificado en:

- 1xx, es un nivel de registro info que indica que no hay error y que el proceso, continua por ejemplo 100 Continue, 101 Switching Protocols…
- 2xx, es un nivel de registro info ok, y significa que se ha realizado la solicitud, por ejemplo 200 OK, 201 Created (para APIs), etc
- 3xx, es una redirección, porejemplo 301 Moved Permanently, 302 Found, etc

Y de los errores puede ser clasificado en:

- 4xx, error de conexión (es de la parte del cliente), por ejemplo, 400 Bad Request, 403 Forbidden, , 404 Not Found,…
- 5xx, error de conexión (en la parte del servidor), por ejemplo, 500 internal server error, 502 Bad Gateway, 503 Service Unavailable, 504 Gateway Timeout, …

Q3.

¿Cuál es la diferencia entre una base de datos relacional y no relacional? Pon dos ejemplos por cada tipo de base de datos.

Las bases de datos relacionales suelen ser todas SQL para su gestión y tienen relaciones entre las tablas, siendo cada tabla un número de filas (registros) y columnas (atributos). Ejemplos son MySQL, PostgreeSQL, Oracle, SnowFlake, etc… Por la parte de BD no relacionales tenemos el resto, es decir, desde bases de datos que almacenan datos de tipo documentos, datos no estructurados, tabulares, sin estructuras fijas, etc. Ejemplos son MongoDB, Cassandra, Neo4j, etc…

Q4.

¿Cuáles son los métodos que permiten saturar un servidor con múltiples peticiones web? Enuméralos y coméntalos brevemente.

- Fuerza bruta, que es una repetición intensiva o incluso masiva de la misma petición, por ejemplo, intentar login en un sitio web, o descargar un archivo repetidamente.
- DDoS, que significa ataque de denegación de servicio, que requiere múltiples dispositivos haciendo solicitudes masivas e intensas a un servidor para llegar a su límite y sobrecargarlo.
- Crawling o Sraping intensivo, si bien el interés no es tumbar el sitio, al realizarse peticiones masivas y excesivas para extraer grandes cantidades de datos, acaba siendo un ataque.

Q5.

Durante el web scrapting, cuando estamos navegando por la estructura anidada de una página web, podemos realizar un análisis vertical o horizontal. Pon ejemplos de cada uno de ellos y comenta la diferencia que existe entre estos dos tipos de análisis.

- Análisis vertical: Navegar en una sección específica, por ejemplo, la zona de descargas de una publicación, inspeccionar los enlaces de descargas.
- Análisis horizontal: es navegar en el mismo nivel y recorrer todas las secciones de ese nivel, por ejemplo, de la zona de publicaciones, visualizar todas ellas, o las secciones del sitio web, etc.
- La diferencia radica en que una pretende tener información profunda de una sección o área y la otra explora diferentes secciones sin profundizar.