Universidad de Alcalá
Escuela Politécnica Superior

Grado en Ingeniería Informática

Trabajo Fin de Grado

Análisis y comparacion de paquetes para el desarrollo de web scraping en R

Autor: David Márquez Mínguez

Tutor: Juan José Cuadrado Gallego

2021

UNIVERSIDAD DE ALCALÁ

ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería Informática

Trabajo Fin de Grado

Análisis y comparacion de paquetes para el desarrollo de web scraping en R

Autor: David Márquez Mínguez

Tutor: Juan José Cuadrado Gallego

Tribunal:

Presidente: . . . . . .

Vocal 1º: . . . . . .

Vocal 2º: . . . . . .

Fecha de depósito: . . . . de . . . . . de . . . . .

Agradecimientos

A todos los que la presente vieren y entendieren.

Inicio de las Leyes Orgánicas. Juan Carlos I

Aqui va la parte de agradecimientos.....

Resumen

Resumen..... correo de contacto: David Márquez Mínguez <david.marquez@edu.uah.es>.

Palabras clave: Trabajo fin de /grado, LaTeX, soporte de español e inglés, hasta cinco....

Abstract

Abstract..... contact email: David Márquez Mínguez <david.marquez@edu.uah.es>.

Keywords: Bachelor final project , LaTeX, English/Spanish support, maximum of five....

Resumen extendido

Con un máximo de cuatro o cinco páginas. Se supone que sólo está definido como obligatorio para los TFGs y PFCs de UAH.

Índice general

Índice de figuras

Índice de tablas

Índice de listados de código fuente

Índice de algoritmos

Capítulo 1

Introducción y objetivos

1.1 Contexto

La *World Wide Web* o lo que comúnmente se conoce como la web, es la estructura de datos más grande en la actualidad, y continúa creciendo de forma exponencial. Este gran crecimiento se debe a que el proceso de publicación de dicha información se ha ido facilitando con el tiempo.

Tradicionalmente el proceso de inserción y extracción de la información se realizaba a través del copy-paste. Aunque este método en ocasiones pueda ser la única opción, esta es una técnica muy ineficiente y poco productiva, pues provoca que el conjunto final de datos no esté bien estructurado. El web scraping o minado web trata precisamente de eso, de automatizar la extracción y almacenamiento de la información extraída de un sitio web [1].

La forma en la que se extraen datos de internet puede ser muy diversa, aunque comúnmente se emplea el protocolo HTTP, existen otras formas de extraer datos de una web de forma automática [2]. Este proyecto, se centra en la metodología existente de obtención de información, de como se tratan los datos y la forma en la que se almacenan. Durante los siguientes apartados se realizará una especificación mas concreta del objetivo del proyecto, así como de la estructura y motivación del mismo.

1.2 Motivación

El proceso de extracción y recopilación de datos no estructurados en la web es un área interesante en muchos contextos, ya sea para uso científico o personal. En ciencia por ejemplo, los conjuntos de datos se comparten y utilizan por múltiples investigadores, y a menudo también son compartidos públicamente. Dichos conjuntos de datos se proporcionan a través de una API [1] estructurada, pero puede suceder que solo sea posible acceder a ellos a través de formularios de búsqueda y documentos HTML. En el uso personal también ha crecido a medida que han comenzado a surgir servicios que proporcionan a los usuarios herramientas para combinar información de diferentes sitios web en su propias colección de páginas.

Además de ser un ambito interesante, el minado web también es un area muy requerida, algunos de las campos de mayor demanda tienen relacion con la venta minorista, mercado de valores, análisis de las redes sociales, investigaciónes biomédicas, psicológia...

[1] Conjunto de subrutinas, funciones y procedimientos que ofrece cierta biblioteca para ser utilizada por otro software como una capa de abstracción. [3]

1.3 Objetivo y limitaciones

Existen muchos tipos de técnicas y herramientas para realizar web scraping, desde programas con interfaz gráfica, hasta bibliotecas software de desarrollo. El objetivo de esta tesis es realizar un análisis cuantitativo de las diferentes técnicas y paquetes software para el desarrollo de web scraping en R.

¿Cuál es la solución más rentable para el minado web? ¿Cuál de las soluciones tiene un mejor rendimiento? Para responder a esta pregunta, se realizará un estudio comparando las diferentes características de los paquetes software, con el objetivo finalmente de poder determinar cuál es el más óptimo en términos de memoria, rendimiento...

En cuanto a las restricciones para el funcionamiento de un scraper, estas pueden ser varias, ya sean legales o por la incapacidad de acceder a una gran parte del contenido no indexado en internet. Aunque el uso de los scrapers está generalmente permitido, en algunos países como en Estados Unidos, las cortes en múltiples ocasiones han reconocido que ciertos usos no deberían estar autorizados [1]. El desarrollo de este proyecto no se verá perjudicado por este tipo de cuestiones, pues solo se limitará al estudio y análisis de los mismos.

1.4 Estructura del documento

Para poder facilitar la composición de la memoria se detalla a continuación la estructura de la misma:

1. **Bloque I: Introducción.**
   - **Capítulo 1: Introducción.**

     En la introducción se especifica tanto el contexto como la motivación a realizar el proyecto, así como las limitaciones esperadas durante la realización del mismo.
2. **Bloque II: Marco teórico.**
   - **Capítulo 2: Web scraping, extracción de datos en la web.**

     Durante este capítulo se explica en que consiste el web scraping, sus posibilidades prácticas y aspectos más generales.
   - **Capítulo 3: Introducción a los paquetes seleccionados y proceso de búsqueda.**

     Se realiza la selección de paquetes y se dictamina cuál ha sido la razón por la que los paquetes han sido seleccionados. Además, se especifican las características principales de cada uno, así como una visión general de sus funcionalidades.
3. **Bloque III: Marco práctico.**
   - **Capítulo 4: Selección de variables de análisis y proceso de estudio.**

     Durante este capítulo se especifica el proceso de análisis a realizar, cuáles son los test a los que los paquetes serán sometidos y que variables se tomarán a estudio para los mismos.
   - **Capítulo 5: Análisis y comparativa de paquetes.**

     Una vez introducidos todos los paquetes y el estudio al que van a ser sometidos, se realizá la comparativa de los mismos. Inicialmente, los paquetes serán analizados uno por uno y finalmente se hará una comparativa con los datos obtenidos.
4. **Bloque IV: Conclusiones y futuras líneas de trabajo.**

Capítulo 2

Web scraping, extracción de datos en la web: Marco teórico

Durante capítulo 1 ya se ha visto de forma genérica cuál es el objetivo del web scraping, así como ciertas limitaciones del mismo. A continuación se explicará que es realmente, cuál es su funcionamiento y sus posibilidades prácticas.

2.1 En que consiste el web scraping

2.2 ¿Como funciona el web scraping?

Capítulo 3

Presupuesto

Blah, blah, blah.

Bibliografía

[1] Wikipedia, "Web scraping," https://es.wikipedia.org/wiki/Web_scraping. [Ultimo acceso 18/octubre/2021].

[2] B. Zhao, *Web Scraping*, 05 2017, pp. 1–3.

[3] Wikipedia, "Interfaz de programación de aplicaciones," https://en.wikipedia.org/wiki/API. [Ultimo acceso 18/octubre/2021].

Universidad de Alcalá
Escuela Politécnica Superior

Universidad
de Alcalá