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

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

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 al scraping en la web: Marco teórico

1.1 ¿Que es realmente el web scraping?

La web es la estructura de datos mas grande en la actualidad, una gran cantidad de datos es almacenada en algun lugar esperando a ser consultada. Tradicionalmente la extracción de dicha información se ha realizado a traves del copy-paste, aunque en ocasiones pueda ser la unica manera, esta es una tecnica muy poco productiva e ineficiente de obtener información. Hasta ahora, no ha existido un termino oficial de la palabra, pero podemos definir *web scraping* o como *"técnica utilizada mediante programas de software para extraer información de sitios web."* [1]

[En esencia, el web scraping se utiliza para obtener datos no estructurados de páginas web y transfórmelo en una presentación estructurada o para almacenarlo en una base de datos externa. También se considera una técnica eficiente para recopilar macrodatos, donde la recopilación de grandes cantidades de datos es importante.]

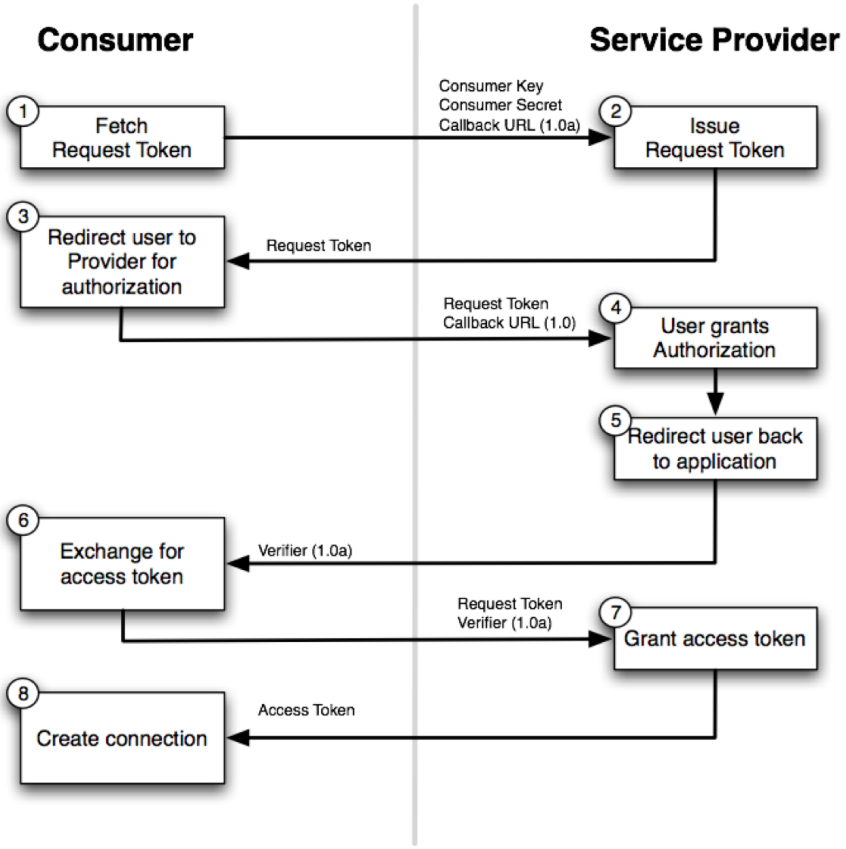

Figura 1.1: Funcionamiento de Oauth1a

Capítulo 2

Presupuesto

Blah, blah, blah.

Bibliografía

[1] Wikipedia, “Web scraping,” https://es.wikipedia.org/wiki/Web_scraping. [Ultimo acceso 14/octubre/2021].

Apéndice A

Funciones implementadas

A.1 Función de extracción de Tweets

Listado A.1: Extracción de Tweets empleando una funcion en R

```
extraccion_tweets <- function(usuario, maxtweets = 100, archivoSalida= NULL){

  #Se crea el nombre de archivo por defecto
  if(is.null(archivoSalida)){
        archivoSalida <- paste0("datos_tweets_", usuario, ".csv")
  }

  #Se comprueba si el archivo csv existe o no
  if(!(archivoSalida %in% list.files())){
    datos_new <- searchTwitter(usuario, n = maxtweets, exclude:retweets)
    datos_new_df <- twListToDF(datos_new)
    write.csv(datos_new_df, archivoSalida)
  }else{
    #Obtengo los datos antiguos
    datos_old <- read.csv(file = archivoSalida)

    #Calculo el id del nuevo tweet a obtener
    ultimo_id <- tail(datos_old, 1)["id"] %>% pull()
    ultimo_id = ultimo_id + 1

    datos_old <- map_if(.x = datos_old, .p = is.numeric, .f = as.character)

    datos_new <- searchTwitter(usuario, n = maxtweets, maxID = ultimo_id, exclude:retweets)
    datos_new <- map_if(.x = datos_new, .p = is.numeric, .f = as.character)

    datos_new_df <- as.data.frame(datos_new)
    datos_old_df <- as.data.frame(datos_old)

    #Se concatenan los nuevos tweets con los antiguos
    datos_concatenados <- bind_rows(datos_old_df, datos_new_df)

    write.csv(datos_concatenados, archivoSalida)
  }
}
```

El objetivo fundamental de esta función es extraer los tweets publicados por un usuario y almacenarlos en un archivo csv. En caso de que exista un archivo con el mismo nombre, se lee y se concatena el nuevo contenido con el antiguo. Los argumentos de entrada, son los siguientes:

1. usuario: reprsenta el identificador del usuario de Twitter.
2. maxtweets: cantidad de tweets que se van a recuperar.
3. archivoSalida: nombre del fichero de salida.

Se debe tener cuidado de no recuperar tweets repetidos, para ello se ha creado una variable que almacena el id del ultimo tweet. Con esto cada vez que se quieran recuperar nuevos tweets, se aumentará en uno dicha variable y se procederá como hasta ahora.

Universidad de Alcalá
Escuela Politécnica Superior

Universidad
de Alcalá