WEB-SCRAPPING - PYTHON

PROYECTO DE COMPUTACIÓN II

DESARROLLO DE UN SCRAPER WEB

ADRIÁN GALLEGO, CARLOS CASTRO, ALEJANDRO ABAD

ÍNDICE

Contenido

INTRODUCCIÓN

OBJETIVO

El objetivo de esta actividad es programar un sistema de web scrapping que recopile de manera Automática información sobre casos de violencia de género que estén publicados en una o varias páginas web.

ENUNCIADO

Paso 1: Buscar y seleccionar una o varias páginas web que recopilen información sobre noticias de violencia de género. Pueden ser páginas de observatorios de violencia, periódicos, informes, etc.

Paso 2: Utilizando el lenguaje de programación python, crear un programa que descargue el html de las webs del paso 1, y haga un parseo para extraer información sobre casos de violencia de género. La información debe ser lo más completa posible y lo más parecida posible a la que se recoge con la interfaz desarrollada en Delphi en la actividad anterior.

PASO 3: Almacenar la información parseada. Se puede guardar en formato texto (CSV, json,,etc...) para que luego se pueda insertar en la base de datos, o podéis hacer que se inserte directamente en la BD desde el programa en Python.

DESARROLLO

Para el desarrollo de esta entrega de PC2 hemos desarrollado un web scraper que se encarga de extraer información relevante sobre casos de violencia de género en España desde 1999 hasta los últimos casos de 2018 y se va actualizando con los casos que nos van aconteciendo a día de hoy. En nuestro caso hemos optado por la página web de la "Plataforma Cordobesa contra la violencia de género" ya que se trata de una web con una gran cantidad de datos para poblar de forma adecuada nuestra base de datos y porque sigue una estructura muy sencilla de cara a la extracción de los datos que requerimos.

Link: http://www.plataformacordobesa.com/computo/?a=

Parte 1

En un primer lugar exploramos múltiples opciones sobre los distintos lugares donde podíamos aplicar web scraping, como la página "feminicidio.net", pero nos resultó demasiado complejo la extracción de datos debido a la estructura inconsistente que seguía.

Parte 2

Posteriormente encontramos la página "Plataforma Cordobesa contra la violencia de género" y comenzamos con la extracción de los datos. Empleamos la librería "BeautifulSoup" de bs4 para Python, siendo este lenguaje el utilizado para la realización de la práctica. El código con el que extraemos los datos en cuestión se encuentra en el archivo "webScrapingViolencia.py" en la carpeta PC2, donde además de trabajar con la minería de datos, agrupamos y guardamos dicha información en un archivo .xlsx gracias a la librería "pandas". El archivo resultante tras la ejecución del programa se genera en la siguiente ruta: "./Excel/webscrapingviolencia.xlsx".

Parte 3

Por otro lado, también hemos empleado la herramienta "Open Refine" para limpiar el conjunto de datos ("webscrapingviolencia.xlsx") gracias a la agrupación en clúster y gracias a la herramienta de "Reconcilie Column" con el objetivo de emplear el "Wikidata reconciliation service for OpenRefine" para filtrar los datos y seguir el formato empleado en la Wikipedia.

Link (WRSOR): "https://tools.wmflabs.org/openrefine-wikidata/"

Generando un segundo excel : "Violencia1999_2018OpenRefine.xls"

Parte 4

En esta última fase del desarrollo de esta práctica empleamos la librería "sqlite3" para insertar datos en nuestra base de datos. Actualmente lo único que insertamos en la base de datos son una serie de URLs que extraemos del periódico "El Pais", pero nuestra idea principal y que se encuentra en desarrollo trata de insertar las URLs que actualmente extraemos del repositorio de noticias de violencia de género de la "Plataforma Cordobesa contra la violencia de género". Aun así, te adjuntamos el código con la inserción a la base de datos en el archivo "Principal.py" con el objetivo de expandir nuestro volumen de datos en la BBDD.

Futuras mejoras

- Lograr insertar en la BBDD todos los datos extraídos de la web que hemos seleccionado.
    - En la tabla "Noticias" pretendemos generar una noticia por cada columna del Excel guardando una *id noticia* única y una URL en caso de que la tuviera.
    - En la tabla "Localización" guardaremos *fecha*, *id comunidad*, *id provincia* e *id municipio* si la tuviera.
    - En la "victima" se guardarán *las iniciales o el nombre*, *el sexo* (Todas serán femeninas) y *la edad* en caso de que fueran conocidas.
    - En el "agresor" se guardarán *las iniciales* o *el nombre*, *el sexo* (Todos serán masculinos) y *la edad* en caso de que fueran conocidas.
    - En "información" se guardará la relación entre víctima y agresor.