Web scraping

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS – Práctica 1

Almudena Caballero Manzanas
Ángel A. Urbina Sánchez
11-4-2022

ÍNDICE

DESCRIPCIÓN

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar herramientas de extracción de datos.

1.- Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información

Nos planteamos recopilar y almacenar, para posteriormente analizar, la información relativa a las posibles **ofertas de estudios de máster**. El objetivo de obtener esta información es el disponer de un único fichero, sin necesidad de visitar cada una de las webs, con la información que consideramos más relevante para elegir el máster considerado.

Elegimos **Emagister** (https://www.emagister.com), web buscadora de másteres, y tratamos de obtener toda la información que nos ayude a la elección del máster deseado. Consideramos que la **información** relevante para desarrollar esta decisión es:

- Descripción del máster: tipología, duración, metodología, …
- Precio
- Entidad
- Programa

2.- Título. Definir un título que sea descriptivo para el dataset

Empleamos como título del dataset *InfoMaster 07-04-22.csv*

3.- Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido

El dataset recoge la información relativa a:

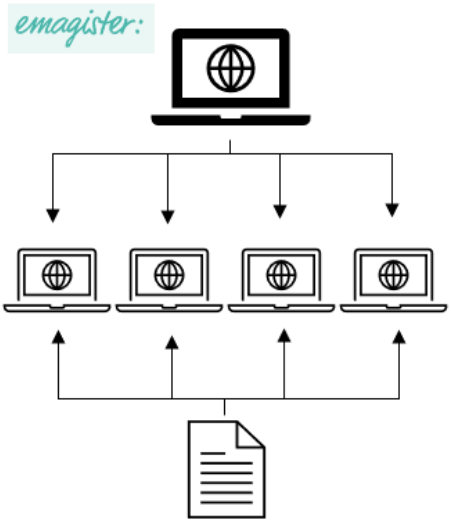
a) Descripción general: descripción relativa a las características del máster impartido por cada entidad como metodología, duración, tipología, fecha de inicio, …
b) Precio, así como si existe o no la posibilidad de financiación
c) Existencia de bolsa de empleo
d) Descripción detallada del máster
e) Programa académico

Nótese que no todas las entidades ofrecen la misma cantidad de información por lo que puede haber algunos valores que, para determinadas entidades, no podamos obtener.

4.- Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido

Hacemos la búsqueda inicial en Emagister, esta búsqueda nos proporciona nuevas webs con la información de cada entidad que imparte el máster buscado. Es la información contenida en cada una de esas webs la que recoge nuestro dataset final.

Conceptualmente, podemos obtener una imagen más clara de cómo hemos obtenido los datos que contiene nuestro dataset en el siguiente esquema:

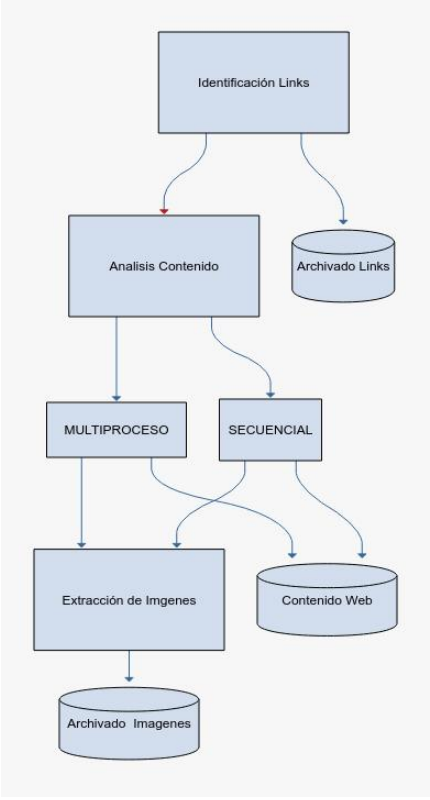


Técnicamente, el proceso consiste en extraer los links de los diferentes másteres de Emagister. Estos links son almacenados en un archivo que posteriormente usaremos para extraer la información de cada uno de ellos. Análogamente, implementamos un proceso de captación de los logos de cada una de las entidades que imparten el máster.

Hemos desarrollado dos maneras diferentes de obtención de la información de los másteres:

- MODO SECUENCIAL El acceso a los diferentes links WEB se desarrolla de forma secuencial (Mayor tiempo de ejecución)
- MODO MULTIPROCESO El acceso a los diferentes links WEB se desarrolla de forma paralela (Menor tiempo de ejecución)

El siguiente esquema sintetiza el proceso explicado anteriormente:



5.- Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido

El contenido del dataset ha sido obtenido mediante técnicas de web scraping, empleando Python, con fecha 07/04/2022. El dataset se compone de las siguientes variables:

i. Nombre: nombre del máster
ii. Entidad: centro educativo que imparte el máster
iii. Precio
iv. Tipología
v. Metodología: online o presencial
vi. Duración
vii. BolsaEmp: indica si el máster dispone o no de bolsa de empleo
viii. Teléfono: teléfono de contacto de la entidad
ix. Descr: descripción detallada de la información relevante del máster
x. Programa: programa académico del máster
xi. UrlImagen: URL del logo de la entidad que imparte el máster
xii. NombreImagen: tipo de imagen (jpg, png, …)
xiii. Web: enlace a la web con la información recopilada

6.- Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto

Emagister, tal y como ellos se definen, es el punto de encuentro entre los que buscan y ofrecen información. Tras más de una década trabajando para hacer de su directorio el más completo del mundo, tanto en volumen como en profundidad de información, cuentan con más de 100.000 centros de formación.

Los principios por los que se rige Emagister son:

- ✓ Hacer accesible la formación a todo el mundo apostando por la creación de un directorio de formación que dé cabida a toda la formación existente (Grados, postgrados, másteres, cursos de especialización, ...) y facilite el poder compartir el conocimiento
- ✓ "Lifelong Learning": consideran que el aprendizaje es un proceso continuo a lo largo de la vida y en todos los ámbitos. Por tanto, debe ser modular, on-demand y hecho a la medida de cada persona

Respecto a los **principios éticos y legales**, hemos actuado de acuerdo con lo marcado según Emagister ya que en sus términos de titularidad y propiedad intelectual e industrial se recoge:

"*El Usuario se compromete a respetar los derechos de Propiedad Intelectual e Industrial de titularidad de EMAGISTER. Podrá* ***visualizar los elementos de las diferentes websites e incluso imprimirlos, copiarlos y almacenarlos en el disco duro de su ordenador o en cualquier otro soporte físico siempre y cuando sea, única y exclusivamente, para su uso personal y privado****. El Usuario deberá abstenerse de suprimir, alterar, eludir o manipular cualquier dispositivo de protección o sistema de seguridad que estuviera instalado en las páginas de EMAGISTER*"

Es por ello por lo que asumimos que el uso moderado de web scraping es adecuado.

7.- Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6

Motivados por la reciente tarea de **elección de un máster**, nos planteamos una forma más cómoda y sencilla de recolectar la información de las distintas webs.

Muchas veces, las tareas de búsqueda se convierten en un proceso tedioso en el que, finalmente, acabas con numerosas ventanas abiertas en las que pierdes la visión de toda la información. Así, pretendemos obtener, de forma rápida y centralizada, una herramienta que nos permita disponer de toda la información útil sobre un máster de interés.

Este objetivo lo conseguimos con Emagister, ya que se trata de una página de búsqueda de formaciones, mediante la cual podemos recopilar la información. Somos, además, partidarios de los principios sobre los que se rigen: información accesible para todos y aprendizaje continuo a lo largo de todos los ámbitos de la vida.

8.- Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección

Para no incurrir en los términos de titularidad y propiedad intelectual e industrial marcados por Emagister, le asignamos a nuestro dataset la licencia CC BY-NC-ND 4.0 ya que ésta permite usar una obra mientras cites al autor, sea para proyectos no comerciales y no se modifique de ninguna manera

CONTRIBUCIONES

Contribuciones	Firma
Investigación previa	ACM, AUS
Redacción de las respuestas	ACM, AUS
Desarrollo del código	ACM, AUS

REFERENCIAS

- SUBIRATS MATÉ, Laia y CALVO GONZÁLEZ, Mireia. *Web scraping* [en línea]. Barcelona: UOC. Disponible en: https://materials.campus.uoc.edu/daisy/Materials/PID_00256970/pdf/PID_00256970.pdf

- Genbeta - Las licencias Creative Commons, explicadas para dummies [en línea]. Disponible en: https://www.genbeta.com/herramientas/licencias-creative-commons-explicadas-para-dummies#:~:text=Atribuci%C3%B3n%2DNoComercial%2DNoDerivadas%20(CC,BY%2DNC%2DND%204.0)&text=Puedes%20usar%20una%20obra%20mientras,por%20ejemplo%2C%20tienen%20estas%20licencias

- https://stackoverflow.com/questions/66876071/extracting-a-complex-substring-using-regex-with-data-from-a-string-in-python

- https://stackoverflow.com/questions/14473180/regex-to-get-a-filename-from-a-url

- https://medium.com/@kunal.rustagi/boost-your-web-crawler-using-multiple-processes-in-python-3cc3ff519226

- https://coderzcolumn.com/tutorials/python/logging-config-simple-guide-to-configure-loggers-from-dictionary-and-config-files-in-python