Publicaciones científicas y libros de ciencia y tecnología

2 de abril de 2021

Índice

1. Presentación y objetivos

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

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.
- Saber identificar los datos relevantes que su tratamiento aportan valor a una empresa y la identificación de nuevos proyectos analíticos.
- Saber identificar los datos relevantes para realizar un proyecto analítico.
- Capturar datos de diferentes fuentes de datos (tales como redes sociales, web de datos o repositorios) y mediante diferentes mecanismos (tales como queries, API y scraping).
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar la capacidad de búsqueda, gestión y uso de la información y de los recursos.

2. Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de web scraping.

3. Desarrollo

1. Contexto

Cada día más personas están más interesadas en la lectura mediante contenido digital, es por esto por lo que la industria del eBook ha permitido ampliar los mercados. La industria de la investigación y la ciencia también están migrando a estas plataformas provocando una mayor difusión y cercanía con la población a temas que antes solo estaban en boca de la comunidad científica.

En el contexto de la actual crisis sanitaria global se ha seleccionado este portal para recolectar datos de las publicaciones que se han estado realizando en la disciplina de la tecnología aplicadas a diferentes ramas áreas, como la salud, educación, conectividad, desarrollo social, entre otros.

Este portal cuenta con una biblioteca de recursos científicos, digitales mayormente, de primera, llegando en ciertos casos a estar disponibles gratuitamente al público en general. Considerando el historial tecnológico de los participantes en este proyecto, se ha limitado el alcance del estudio a recolectar información de los libros que se han categorizado como ciencias de la computación (Computer Science) desde el 2015 hasta el presente año.

Existe la necesidad de conocer los temas donde la comunidad científica ha sido más prolífica, lo cual permitirá conocer las áreas donde se pueden implementar nuevas soluciones a las problemáticas de la sociedad. Así mismo, con este proyecto se pretende conocer los ámbitos donde aún los científicos deben invertir sus esfuerzos para investigar y publicar material digital de calidad.

Por otra parte, la tendencia hacia una mayor especialización en los sectores industriales ha incrementado el número de estudiantes de másteres y doctorados en el país por casi un 11% (datos del 2018), lo que ha conllevado a una mayor difusión de sitios web como “springer” ya que son fuentes de contenido refutadas y de fácil acceso.

Por este nuevo interés y avance de la forma de comercializar este contenido, nace una necesidad de conocer más acerca de la oferta y las opciones más convenientes a la hora de adquirir este tipo de contenido, además de conocer cuáles son las tendencias mediante el número de publicaciones realizadas por año.

2. Título del dataset

Dataset de libros electrónicos de ciencias de la computación publicados en Springer desde el 2015 al 2021.

3. Descripción del dataset

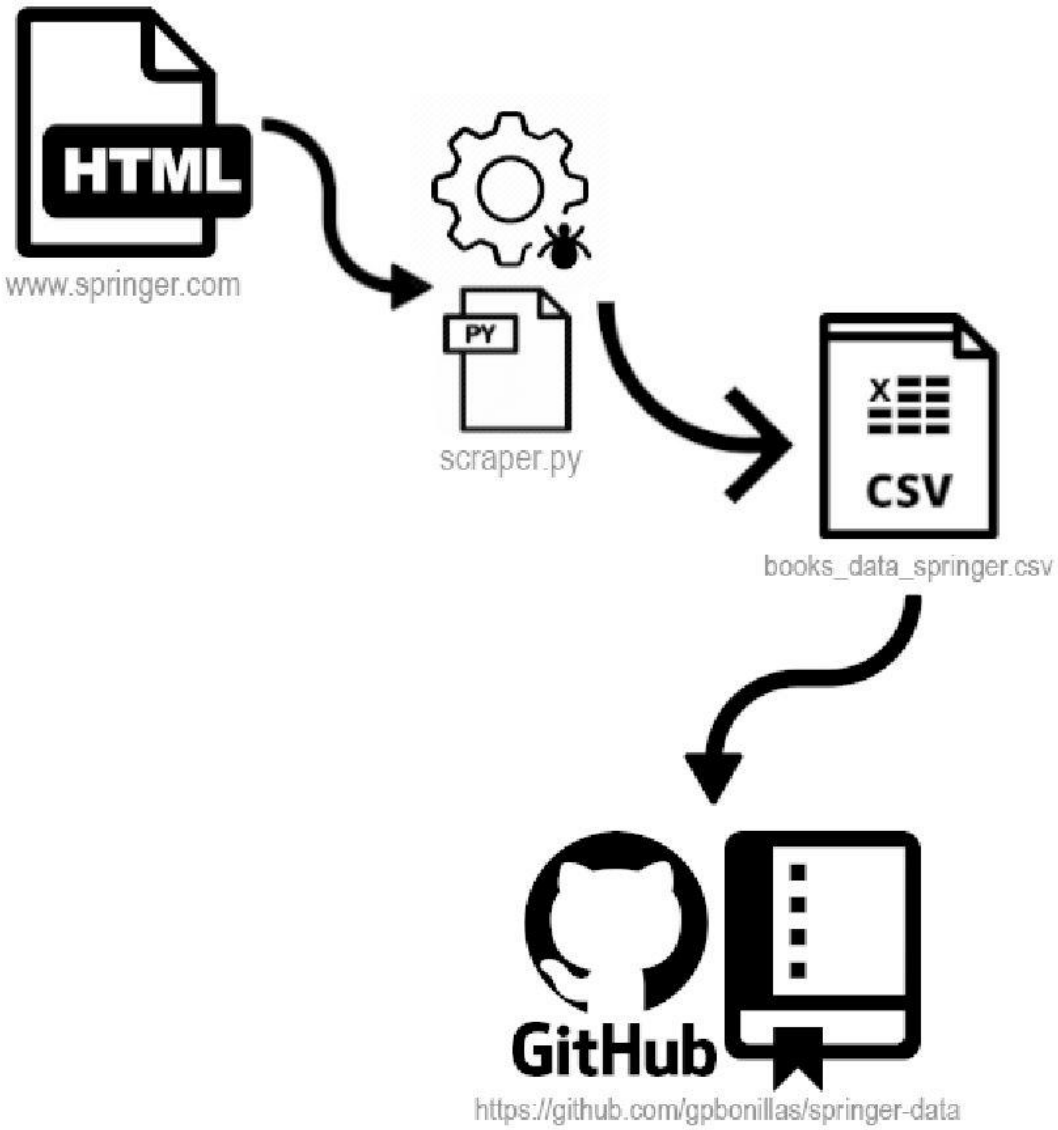
Se ha seleccionado el sitio web de Springer, ya que ofrece una gama muy amplia libros en formato físico y digital de ciencia, tecnología, medicina y más.

Considerando que en este sitio web ofrece la posibilidad de adquirir libros en formato físico y digital se ha decidido analizar cuál es la diferencia de precios entre estos dos formatos he intentar conocer cuando es más conveniente comprarlos en un determinado formato. Además, se quiere analizar la frecuencia de publicaciones anuales en las diferentes temáticas que corresponden a la disciplina de ciencias de la computación (Computer Science), desde el año 2015 a la actualidad.

El dataset está basado en poder encontrar toda la información sobre un libro publicado en el sitio Springer, partiendo desde el nombre hasta el costo; con las siguientes características:

- **Formato:** CSV
- **Fuente de datos:** https://www.springer.com
- **Nombre del dataset:** books_data_springer.csv
- **Número de páginas consultadas:** 500.
- **Primera línea:** Contiene las etiquetas o encabezados de los campos.
- **Separador de los campos**: coma (,)
- **Número de variables o campos**: 15.
- **Número de filas (aprox**.): 10000

4. Representación gráfica



5. Contenido

Como hemos mencionado anteriormente, el periodo del tiempo considerado para extraer datos está entre el 2015 y el 2021. La recolección se ha aplicado únicamente sobre la disciplina Ciencias de la Computación (Computer Science)

El dataset contiene toda la información relevante sobre un libro publicado en Springer. Se han considerado los siguientes campos a extraer:

Campos del dataset:

Nombre del campo	Tipo	Descripción	Ejemplo
Page	Numérico	Número de la página (en Springer) desde donde se obtuvo la información	258
title	Texto	Título o nombre del libro	'Java 9 Revealed'
topic	Texto	Temática central del libro	'Java'
isbn	Texto	Identificador ISBN del libro	'978-1-4842-2592-9'
pages	Texto	Número de páginas del libro. Algunos contienen la numeración romana inicial.	'XXIV, 521'
year	Numérico	Año del copyright del libro	2017
online_access	Numérico	Código numérico que indica acceso online.	0 = NO, 1 = SI
format	Numérico	Código numérico que indica el formato del libro	1 = ebook, 2 = hardcover, 3 = online, 4 = softcover, 5 = 2 formatos, 6 = más de 2 formatos.
editorial	Texto	Casa Editorial o editora del libro	'Apress'
ebook-price	Numérico	Precio del libro en formato ebook	29.99
hardcover-price	Numérico	Precio del libro en formato hardcover (pasta dura)	29.99
softcover-price	Numérico	Precio del libro en formato softcover (pasta blanda)	49.99
print-price	Numérico	Precio del libro en formato impreso (print)	49.99
print-ebook-price	Numérico	Precio del libro en formato print-ebook	1,199.99
authors	Texto	Autor (es) del libro.	"Zhang, Y. (Ed) (2021)"

Se han extraído campos o variables desde la página donde se encuentran el listado de libros, cuyo número se indica en el dataset, y también de la página donde

reposa la información bibliográfica de cada uno.

Se ha ejecutado el script para extraer datos en rangos de 100 páginas con un pequeño retraso entre 3 y 5 segundos para evitar posibles bloqueos por parte del administrador del portal de Springer. Cabe mencionar que durante la extracción del dataset no se tuvo problemas con bloqueos por parte del administrador del portal.

En el caso de las variables (campos) donde no se ha encontrado datos por diferentes razones se ha definido valores de “NA” o nulos (vacíos).

6. Agradecimientos

Agradecemos a la web de Springer, la cual consideramos de gran valor, ya que es que una de las más completas del ámbito científico con miles de libros publicados a la venta lo que nos ha ayudado a poder generar el dataset que responde a los cuestionamientos ya mencionados.

También agradecemos a los portales GitHub y Zenodo, que han proporcionado un medio para cargar nuestro proyecto y dataset, respectivamente; con esto ha sido posible el acceso público para los fines que la comunidad considere oportunos.

Para le extracción de la información hemos usado herramientas de código libre, como es, el lenguaje Python con la librería *BeautifulSoup*, por lo cual también nuestros agradecimientos van dirigidos a toda la comunidad de desarrolladores que mantienen y dan soporte a estas tecnologías.

Se detalla el propietario de la página web de Springer, según los datos obtenidos por el script *owner.py*, que se encuentra en el repositorio:


```
{
  "domain_name": [
    "SPRINGER.COM",
    "springer.com"
  ],
  "registrar": "Eurodns S.A.",
  "whois_server": "whois.eurodns.com",
  "referral_url": null,
  "updated_date": [
    "2020-05-22 03:24:22",
    "2020-05-22 05:36:31"
  ],
  "creation_date": [
```

```
      "1997-05-29 04:00:00",
      "2009-05-28 00:00:00"
    ],
    "expiration_date": [
      "2021-05-28 04:00:00",
      "2021-05-27 00:00:00"
    ],
    "name_servers": [
      "PDNS1.ULTRADNS.NET",
      "PDNS2.ULTRADNS.NET",
      "PDNS3.ULTRADNS.ORG",
      "PDNS4.ULTRADNS.ORG",
      "PDNS5.ULTRADNS.INFO",
      "PDNS6.ULTRADNS.CO.UK",
      "pdns1.ultradns.net",
      "pdns2.ultradns.net",
      "pdns3.ultradns.org",
      "pdns4.ultradns.org",
      "pdns5.ultradns.info",
      "pdns6.ultradns.co.uk"
    ],
    "status": [
      "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
      "clientTransferProhibited http://www.icann.org/epp#clientTransferProhibited"
    ],
    "emails": [
      "legal@eurodns.com",
      "legalservices@eurodns.com",
      "hostmaster@springernature.com"
    ],
    "dnssec": "unsigned",
    "name": "Schipper Jaap",
    "org": "Springer Nature B.V.",
    "address": "van Godewijckstraat 30",
    "city": "Dordrecht",
    "state": null,
    "zipcode": "3311 GX",
    "country": "NL"
  }
```


Nuestro proyecto no tiene como referencia o se basa en estudios previos, por tal

motivo creemos que en el futuro podrá ser usado como guía para otros interesados en la temática, permitiendo realizar estudios más completos e integrales sobre los datos de libros publicados en el portal de Springer.

Proyectos similares al nuestro lo podemos encontrar en el portal Kaggle, donde se encuentra publicado un dataset[1] con un listado completo de libros del portal goodreads.com.

7. Inspiración

Hemos elegido estos datos porque responde fácilmente en qué formato de libro es más conveniente comprar de acuerdo con los precios.

Otras preguntas que se podrían responder con este dataset son:

- Frecuencia anual de libros o revistas publicadas en los tópicos que corresponden a la disciplina de Ciencias de la Computación que se han publicado entre 2015 y 2021.
- Conocer las temáticas de investigación donde el costo de adquirir libros es más elevado o más bajo.
- Conocer la disponibilidad de formatos entre los libros y revistas del portal.
- Conocer las temáticas que mayor material bibliográfico con acceso en línea tiene.

Por otra parte, este conjunto de datos nos permite introducirnos en el mundo del scraping ya que su contenido es estructurado, estable y diverso. Con estas características se puede realizar un trabajo de mayor confiabilidad lo que conlleva a análisis más precisos, en nuestro caso poder conocer en qué formatos es más conveniente adquirir un libro o a que tópicos se ha centrado la comunidad científica desde el 2015 a la fecha.

El dataset también podría ser usado como una fuente alternativa en la búsqueda de contenido por autores o tópicos ayudando en la difusión de este tipo de contenido.

Aunque el alcance de este proyecto no abarca extraer información de libros de las diferentes disciplinas (como astronomía, negocios y administración, química, ciencias de la tierra, economía, salud pública, entre otros) que tiene el portal, para futuros estudios se podría considerar incluir esta variable, lo que sin duda permitirá tener otra perspectiva de los datos recolectados.

1 Kaggle, *Goodreads-books* [en línea]. [Fecha de consulta: 13 de abril de 2021] Disponible en: https://www.kaggle.com/jealousleopard/goodreadsbooks

Existen libros que comprende varias líneas investigativas en diferentes temáticas del saber, por lo que analizar un dataset con estas variables adiciones permitiría obtener un conocimiento más global de cómo se desarrolla la investigación en la comunidad científica.

8. Licencia

Se ha seleccionado la licencia MIT, ya que es una licencia permisiva breve y simple con condiciones que solo requieren la preservación de los avisos de licencia y derechos de autor. Los trabajos con licencia, las modificaciones y los trabajos más grandes pueden distribuirse bajo diferentes términos y sin código fuente.

Permisos

- Uso comercial
- Modificación
- Distribución
- Uso privado

Limitaciones

- Responsabilidad
- Garantía

9. Código fuente

El código fuente y la demás documentación del proyecto se encuentra alojada en el siguiente repositorio: https://github.com/gpbonillas/springer-data. La estructura general del proyecto es la siguiente:

- **src**: Contiene 3 archivos, según las siguientes indicaciones:
  - **owner.json:** Respuesta JSON al momento de ejecutar el script owner.py
  - **owner.py:** Script que hace uso de la librería *python-whois,* lo que permite conocer el propietario de la página.
  - **scraper.py:** Script que contiene toda la lógica de extracción de los datos del portal de Springer.
- **data**: Contiene el dataset luego de haber ejecutado el script scraper.py
- **docs**: Contiene la documentación referente al proyecto.
- **LICENSE**: Contiene la declaración de la licencia usada para este proyecto. En este caso se ha usado la licencia MIT.
- **README.md**: Contiene una breve descripción del proyecto.

Con la finalidad de evitar los riesgos de aplicar web scraping al extraer los datos se han aplicado las siguientes buenas prácticas[2]:

- Demoras en las consultas a la página, con unos cortos retrasos entre cada petición al servidor y,
- Gestión de valores nulos o no existentes se lo ha implementado.
- Uso de headers personalizados en las peticiones al servidor. Principalmente se ha modificado la cabecera *'User-Agent'*.

Para ejecutar los scripts del proyecto es necesario instalar las siguientes librerías:

- `pip install requests`
- `pip install beautifulsoup4`
- `pip install python-whois`

10. Dataset

El dataset ha sido publicado en el portal Zenodo, con una breve descripción. El DOI del dataset publicado es el siguiente: `10.5281/zenodo.4683793`

4. Referencias

Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.

Springer, *About Springer* [en línea]. Actualizada: 2021. [Fecha de consulta: 02 abril de 2021]. Disponible en: https://www.springer.com/la/about-springer

Wikipedia, *Springer Science+Business Media* [en línea]. Actualizada: 2021. [Fecha de consulta: 02 abril de 2021]. Disponible en: https://es.wikipedia.org/wiki/Springer_Science%2BBusiness_Media

Crummy, *Beautiful Soup Documentation* [en línea]. [Fecha de consulta: 10 de abril de 2021] Disponible en: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Europapress, *Crecimiento de la comunidad estudiantil en másteres* [en línea]. Actualizada: 2021. [Fecha de consulta: 13 abril de 2021]. Disponible en: https://www.europapress.es/sociedad/educacion-00468/noticia-numero-estudiantes-master-crecio-112-curso-pasado-superar-190000-universidad-espanola-20180611180218.html

[2] Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC

5. Contribuciones

Contribuciones	Firma
Investigación previa	L.A.T.G., G.P.B.S.
Redacción de las respuestas	L.A.T.G., G.P.B.S.
Desarrollo código	L.A.T.G., G.P.B.S.