**Publicaciones científicas y libros de ciencia y tecnología**

2 de abril de 2021

**Índice**

[1. Presentación y objetivos 3](#_Toc66122939)

[2. Competencias 3](#_Toc66122940)

[3. Desarrollo 4](#_Toc66122941)

[4. Referencias 6](#_Toc66122942)

# 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.

# 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.

# Desarrollo

1. **Contexto**

Cada día más personas están más interesadas en la lectura mediante contenido digital, es por esto 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.

Por otra parte, la tendencia hacia una mayor especialización en los sectores industriales ha incrementado el numero 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 mas convenientes a la hora de adquirir este tipo de contenido, además de conocer cuales son las tendencias mediante el número de publicaciones realizadas por año.

1. **Título del dataset**
2. **Descripción de la web y enfoque**

Se ha seleccionado la el sitio web de springer, ya que ofrece una gama muy amplia libros en formato físico y digital de ciencia, tecnología y medicina dado que es una de las mayores editoriales de libros y la segunda más grande a nivel mundial en publicaciones científicas.

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 adquirirlos en un determinado formato. Por otra parte, también analizar la frecuencia de publicaciones por año en las diferentes temáticas para conocer las tendencias investigativas entre los años 2015 y 2021.

1. **Descripción del dataset**

El dataset se ha enfocado en responder los cuestionamientos del punto anterior, para esto se han acordado las siguientes consideraciones:

* 1. El formato del archivo es en csv dada su fácil manejo y difusión.
  2. El nombre de los campos está en la primera fila.
  3. Se ha delimitado los campos por “,”.
  4. El número de columnas son 15.
  5. Los valores no numéricos están con comillas dobles.
  6. El nombre del archivo es…

1. **Representación gráfica**

Poner unos plot de los campos, puede ser por temas, rango de páginas, costos, formatos.

1. **Contenido del dataset**

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.

Campos del dataset:

**Fec\_Informacion:** Fecha en la cual se ha consultado la información en el sitio web, el formato es dd/mm/yyyy hh:mm:ss

**Nombre\_Libro:** Nombre del libro.

**Autor (es):** Autor (es) del libro.

**Tema(s):** Tema(s) del que se trata el libro.

**Categoría:** Sección en la que aparece el libro dentro de la web.

**ISBN:** Identificador ISBN del libro

**Anio\_Copyright:** Año del copyright del libro.

**Numero\_Paginas:** Número de páginas del libro.

**Precio\_EBook:** Precio en dólares del formato digital.

**Precio\_Hardcover:** Precio en dólares del formato físico.

**Precio\_Softcover:** Precio en dólares del formato tapa blanda.

**Online:** ¿Tiene acceso online?

**Tipo\_Formato:** Identificador si el libro esta solo en formato ebook (1), hardcover (2), online (3), softcover (4), 2 formatos (5) o más de 2 formatos (6).

1. **Agradecimientos**

La información ha sido obtenida desde 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.

Para la extracción de la información hemos usado el lenguaje Python combinado con técnicas de web scraping presentadas en el aula, así como también contendió general de algoritmos en internet.

1. **Inspiración**

Hemos elegido esta temática de libros científicos para ayudar en la difusión de este tipo de sitios webs en la comunidad educativa, así como también conocer las tendencias investigativas que forjan el futuro de la sociedad.

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 caracterisicas se puede realizar un trabajo de mayor confiabilidad lo que conllevar a análisis mas precisos, en nuestro caso poder conocer en que formatos es más conveniente adquirir un libro o a que tópicos se ha centrado la comunidad científica desde el 2015 a la fecha.

El data set 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.

1. **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

# Referencias

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

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