# Práctica 1: Web scraping

Geovanny Risco y Robert Novak

8 de noviembre de 2021

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

   Los autores de esta práctica teníamos un gran interés en realizar una extracción de datos en algún ámbito relacionado con ciencias de la salud. Adicionalmente, buscábamos una página web con un diseño y arquitectura que nos supusiera un reto para asentar de forma más profunda los conocimientos de la asignatura. En este contexto, nos decantamos por la página https://cima.aemps.es/cima/publico/home.html. Esta página es una aplicación que permite realizar consultas bajo distintos criterios sobre la información de los medicamentos. Al hacer búsquedas sobre medicamentos u otros aspectos relacionados con ellos, la página proporciona un listado con todos los medicamentos que cumplen los criterios de búsqueda. Adicionalmente, al seleccionar un medicamento en concreto se proporciona una serie de características sobre ellos heterogénea, no todos tienen las mismas características (para más detalle ver el apartado 5). Esta heterogeneidad en la información para extraer la información en combinación con la temática médica nos ha parecido una combinación adecuada para la elección de la página.

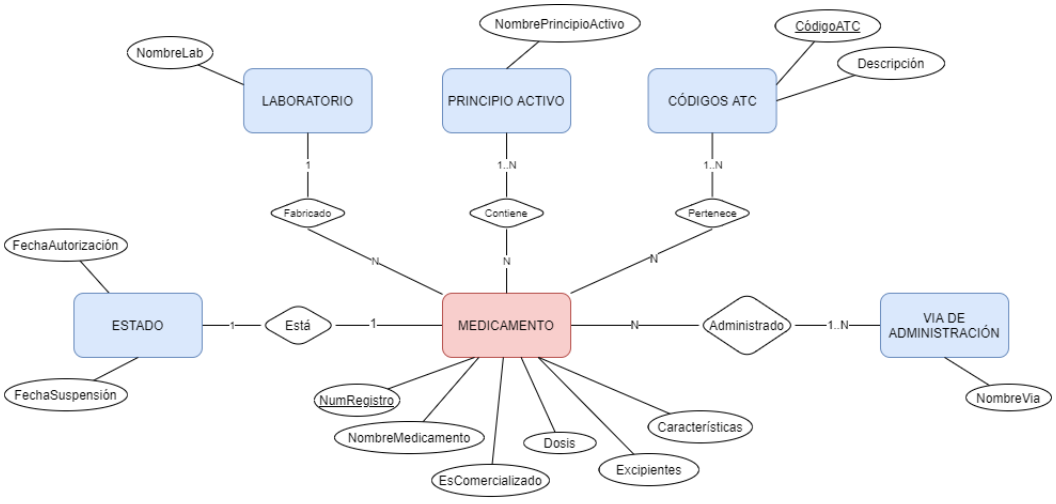
2. **Título. Definir un título que sea descriptivo para el dataset.**

   El título que hemos escogido es Medicamentos registrados por el gobierno de España.

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.**

   La Agencia Española de Medicamentos y Productos Sanitarios (AEMPS), adscrita al ministerio de sanidad del gobierno de España realiza una serie de actividades relacionadas con la autorización, seguimiento y registro de los medicamentos. La información generada por sus actividades, junto a las características presentes en los envases de los medicamentos es lo que se recoge en el conjunto de datos extraído.

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




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

   Los campos que incluye el dataset son los siguientes:

- **Número de registro**: La autorización del medicamento se inscribe de oficio en el Registro de Medicamentos de la Agencia Española de Medicamentos y Productos Sanitarios. Cada número de registro se refiere a una composición, una forma farmacéutica, una dosis por unidad de administración incluyendo todas las presentaciones para la venta. Cada una de las presentaciones es identificada por su correspondiente Código Nacional.
- **Medicamento**: Nombre del medicamento
- **Laboratorio**: Laboratorio donde se ha desarrollado el producto
- **Autorizado**: Indica si el medicamento ha sido autorizado por la AEMPS
- **Fecha autorización**: Fecha en la que se ha autorizado el medicamento
- **Suspendido**: Indica si la AEMPS ha suspendido el medicamento.
- **Fecha Suspensión**: Fecha en la que el medicamento se ha suspendido.
- **Comercializado**: Indica si el medicamento se ha comercializado
- **Vías administración**: Una lista sobre las posibles formas en las que se puede administrar el medicamento
- **Dosis**: Cantidad de medicamento por uso
- **Formas farmacéuticas**: Disposición individualizada a que se adaptan los fármacos (principios activos) y excipientes (materia farmacológicamente inactiva) para constituir un medicamento
- **Principios activos**: Sustancias destinadas a la fabricación del medicamento
- **Excipientes**: Aditivos que se añade al principio activo para darle forma, conservarlo, facilitar su ingesta o regular su actividad en nuestro organismo.
- **Características**: Lista de de algunas características presentes en el medicamento
- **Códigos ATC**: Sistema de clasificación de los medicamentos en función de su lugar de acción, uso terapéutico y estructura.

El conjunto de datos que se presenta en Zenodo se ha recogido el 5 de noviembre de 2021 de la página web descrita en el apartado 1. Los datos que se recogen en el dataset tienen una antigüedad que data de 1927 (primer medicamento autorizado) y va desde entonces hasta la actualidad.

La forma de recoger los datos ha sido mediante el lenguaje de programación Python y, por las características dinámicas de la página web y la necesidad de esperar un tiempo para que cargue el contenido, nos hemos visto obligados a usar la librería de Selenium ya que con BeautifulSoup no era suficiente (https://stackoverflow.com/questions/51213717/beautiful-soup-not-loading-the-entire-page).

Previamente al proceso de la extracción de datos hemos accedido al robots.txt de la página. Mediante el robots hemos comprobado que no se deshabilitan los sitios de la página en los que realizamos web scraping. A continuación, explorando el sitemap se puede ver que la característica más llamativa es la cantidad de páginas con la estructura "https://cima.aemps.es/cima/publico/detalle.html?nregistro=..."donde ... simboliza el número de registro del medicamento. Accediendo a cada una de las páginas se puede apreciar como esa familia de enlaces son los que nos interesa para acceder a la información que queremos scrapear. Haciendo una búsqueda del tipo "site:cima.aemps.es/cima/"nos hacemos una idea de la magnitud de la página. Aparecen 109 000 resultados. No obstante, muchos de ellos son pdfs adjuntos en las distintas secciones de la página.

Finalmente, algunos de los problemas que nos hemos encontrado en la realizacion del script con selenium son que para acceder a la parte de la página donde están todos los elementos del medicamento necesitábamos hacer un scroll hasta el final de la página, pero cada vez que se hacía una iteración se cargaban los elementos de 25 en 25. Como el listado de los resultados no está paginado la complejidad algorítmica está por encima de la lineal. Otro problema al que nos hemos enfrentado es que la página no guardaba en ocasiones los estados internos que se han hecho previamente al navegar por ella. De esta forma, no hemos podido hacer el scrapeo haciendo click en cada elemento y dando el botón de atrás que era la forma más eficiente (midiendo tiempos). Sin embargo, hemos adoptado esta solución para los primeros 25. Para los restantes, hemos ido abriendo un nuevo driver cada vez.

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.**

   Para actuar de acuerdo a los principio éticos y legales accedemos a la sección de avisos legales de la página web https://www.aemps.gob.es/avisoLegal/. Tras realizar una lectura completa, encontramos que el dominio www.aemps.gob.es es de titularidad de la Agencia Española de Medicamentos y Productos Sanitarios, con sede en C/ Campezo 1, 28022, de Madrid. Así, es la AEMPS a la que debemos la posibilidad de la extracción de estos datos, organización a la cual estamos agradecidos. Al ser la AEMPS un organismo público con el poder de realizar la evaluación y autorización de medicamentos no podemos encontrar análisis anteriores que nos proporcione la información que proporciona el dataset. Del mismo modo los datos son públicos y no hay inconveniente en extraerlos.

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.**

   Este conjunto de datos es interesante para cualquier profesional sanitario e, incluso, para usuarios sin muchos conocimientos médicos que requieran de alguna consulta rápida sobre alguna característica de algún medicamento particular. De esta forma, los análisis posibles sobre el dataset pueden ser muy variadas en función de las necesidades y objetivos que se persigan. Por ejemplo, a nivel de paciente, se puede usar para consultar si alguno de los medicamentos de los que se dispone en el hogar se ha suspendido por alguna razón para no usarlo dado el caso o si existe algún principio activo al que se pueda ser alérgico. A nivel científico, se puede usar para contestar preguntas de distinto tipo, por nombrar algunas: ¿Qué laboratorios fabrican medicamentos más exitosos (analizando los medicamentos suspendidos)? ¿Está disponible un cierto medicamento? ¿Cuáles son los principios activos más usados? ¿Qué excipientes son los más usados para los distintos principios activos de los medicamentos?

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

   - **Released Under CC0: Public Domain License.**
   - **Released Under CC BY-NC-SA 4.0 License.**
   - **Released Under CC BY-SA 4.0 License.**
   - **Database released under Open Database License, individual contents under Database Contents License.**
   - **Other (specified above).**
   - **Unknown License.**

   Seleccionamos la licencia Released Under CC0: Public Domain License. Esto es porque la página web de la que nos hemos servido para la extracción de datos es de dominio público. Así,

   - Renunciamos a los derechos de la obra bajo las leyes de derechos autorales en todo el mundo, incluyendo todos los derechos conexos y afines, en la medida permitida por la ley.
   - Se puede copiar, modificar, distribuir e interpretar el conjunto de datos, incluso para propósitos comerciales, sin pedir permiso.
   - No nos hacemos responsables del uso del dataset y, al igual que se indica en los avisos legales de la página, no podemos dar garantía de no contradicción con los datos impresos de la Administración competente debido a la tardanza en las actualizaciones. No nos hacemos responsables tampoco de los perjuicios causados por ello.

9. **Código. Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.**

   El código con el que se ha generado el dataset es python y se encuentra en el siguiente repositorio de GitHub: https://github.com/roberttnovak/MedicineScraper

10. **Dataset. Publicar el dataset obtenido(*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.**

    El enlace del DOI de Zenodo es el siguiente: https://doi.org/10.5281/zenodo.5651781

Enlace de vídeo: https://drive.google.com/file/d/1cW2BVrSITteRmjxXpTRL6hqAY-iHwXvj/view

| Contribuciones | Firma |
|---|---|
| Investigación previa | Geovanny Risco, Robert Novak |
| Redacción de las respuestas | Geovanny Risco, Robert Novak |
| Desarrollo del código | Geovanny Risco, Robert Novak |