



PRA1 - Obtención de un Dataset mediante Webscraping

Tipología y Ciclo de Vida de los Datos

Índice

1	Contexto	1
2	Título	1
3	Descripción del Dataset	1
4	Representación Gráfica	2
5	Contenido	3
6	Agradecimientos	4
7	Inspiración	4
8	Licencia	6
9	Código	6
10	Dataset	6
11	Contribuciones	6

Índice de figuras

1	Esquema gráfico del proceso	2
2	Web del Medicamento	2
3	Salida del CSV	3

1. Contexto

“El Centro de Información online de Medicamentos Autorizados (CIMA) de la AEMPS proporciona a los ciudadanos y profesionales toda la información sobre los medicamentos de forma comprensible para conseguir de esta forma su correcta utilización”¹. De esta forma presentaba el Gobierno, en junio de 2016, la página de Sanidad que contiene la información de múltiples medicamentos.

Hoy en día es fácil encontrar cualquier información en Internet, pero en el contexto de la medicina es muy difícil distinguir lo que proviene de un origen fiable, o incluso qué hacer con una cantidad tan grande de datos. En este sentido, es cierto que el objetivo de la interfaz anterior es dar la posibilidad al usuario (ya sean ciudadanos o profesionales sanitarios) de tener la información completa sobre un medicamento de manera rápida, fácil y accesible, pero en el momento en que lo que se pretende es realizar un análisis más detallado es complicado tratar tantos datos de manera cómoda y útil.

El objetivo de este proyecto es la creación de un software que permita el volcado de información detallada de múltiples medicamentos en un conjunto de datos final, en función de la búsqueda del usuario para poder atender de esta manera sus necesidades de análisis.

2. Título

Dado que el software atiende diferentes necesidades en función del tipo de usuario, el título del conjunto de datos puede variar según la búsqueda inicial.

Un ejemplo de un título obtenido en el proyecto es CINARIZINA.CSV (incluido en el entregable), pero el título puede verse modificado según si la búsqueda es referente a un laboratorio, un principio activo, un medicamento, etc.

3. Descripción del Dataset

El conjunto de datos obtenido mediante el programa desarrollado contiene un catálogo de medicamentos en función del término de búsqueda. Para los diferentes registros, se presentan todas las características presentadas en el detalle del medicamento en la web CIMA comentada en el apartado 1. Esta información hace referencia al número de referencia del medicamento, la fecha de autorización, el laboratorio de producción, los principios activos u otras características, entre otros.

¹<https://www.aemps.gob.es/informa/notasinformativas/laaemps/2016/ni-aemps.06-2016-cima-premio-cuidadania/?lang=en>

4. Representación Gráfica

El flujo del proceso seguiría algo similar a lo que se representa a continuación:

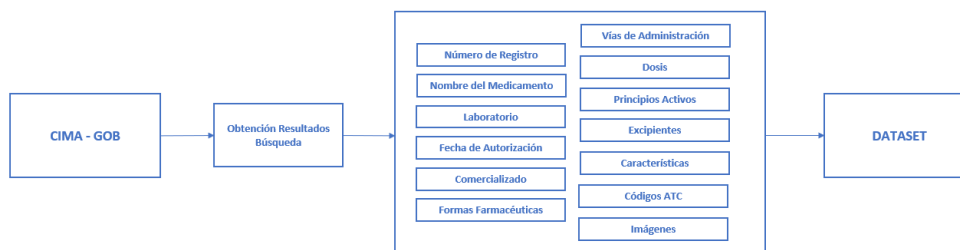
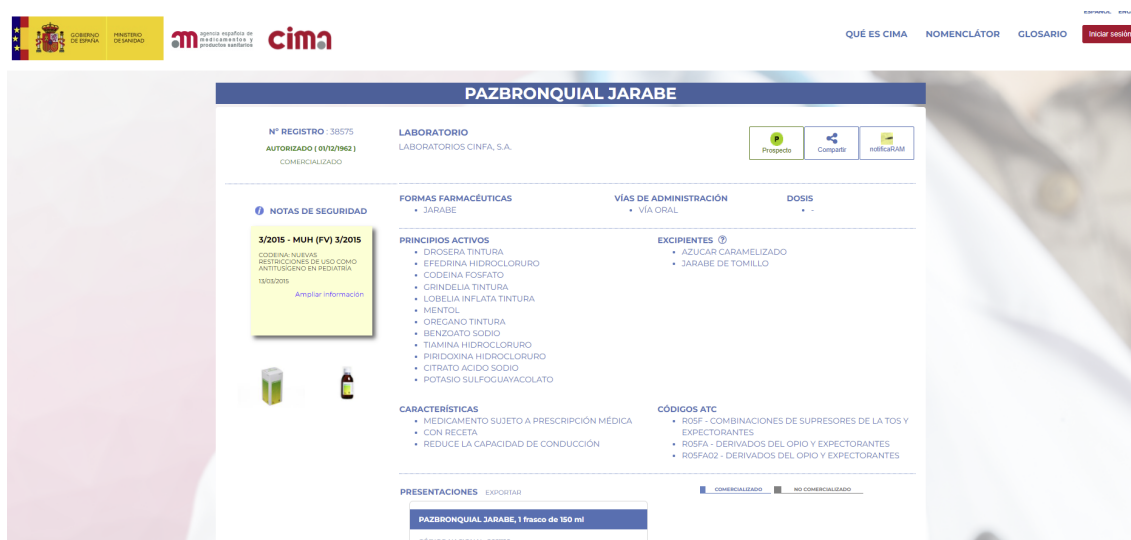


Figura 1: Esquema gráfico del proceso

La información se obtiene de CIMA, que para cada medicamento tiene una presentación del siguiente estilo:



La imagen muestra la interfaz de usuario de la web del medicamento PAZBRONQUIAL JARABE. En la parte superior, se encuentran los logos del Gobierno de España, el Ministerio de Sanidad y el servicio regulatorio de medicamentos, junto al logo de CIMA. A la derecha, hay enlaces para 'QUÉ ES CIMA', 'NOMENCLÁTOR', 'GLOSARIO' y un botón de 'Iniciar sesión'. El contenido principal está dividido en secciones: 'NOTAS DE SEGURIDAD' (con una alerta de 3/2015 - MUH (FV) 3/2015), 'FORMAS FARMACÉUTICAS' (Jarabe), 'VÍAS DE ADMINISTRACIÓN' (Vía oral), 'DOSIS', 'PRINCIPIOS ACTIVOS' (Drosera tintura, efedrina hidrocloreuro, codeína fosfato, grindeña tintura, lobelia inflata tintura, mentol, oregano tintura, benzoato sódico, tiamina hidrocloreuro, piridoxina hidrocloreuro, citrato ácido sódico, potasio sulfocavacolato), 'EXCIPIENTES' (Azúcar caramelizado, jarabe de tomillo), 'CARACTERÍSTICAS' (Medicamento sujeto a prescripción médica, con receta, reduce la capacidad de conducción), 'CÓDIGOS ATC' (R05F - combinaciones de supresores de la tos y expectorantes, R05FA - derivados del opio y expectorantes, R05FA02 - derivados del opio y expectorantes) y 'PRESENTACIONES' (PAZBRONQUIAL JARABE, 1 frasco de 150 ml). En la parte inferior, se indica el código nacional 802728.

Figura 2: Web del Medicamento

El programa recoge toda la información y la vuelca en el dataset que se guarda en la ruta definida.

5. Contenido

El dataset obtenido contiene un total de 13 campos. Estos se representan de la siguiente manera:

A	B	C	D	E	F	G	H	I	J	K	L	M	N
registration num	name	company	authorization date	commercialized	pharmaceutical dose form	routes administration	strength	active ingredients	excipients	characteristics	atc codes	images url	
53105	CUNADIL CAPSULAS	Laboratorio Stada, S.	01/06/1975	VERDADERO	[CÁPSULA DURA]	[VÍA ORAL]	[20.0, 1.0, 5.0]	[CINARIZINA], [DIME]		[MEDICAMENTO SU]	[C04A - VASODILAT]	[https://cima.emps.es/cima/fotos/]	
84721	DIZINEL 20 MG/40 MI	Italfarmaco, S.A.	11/12/2019	VERDADERO	[COMPRIMIDO]	[VÍA ORAL]	[20.0, 40.0]	[CINARIZINA], [DIME]	[CROSCARMELOSA]	[MEDICAMENTO SU]	[N07C - PREPARAD]		
53025	STUGERON CAPSULA	Esteve Pharmaceuti	01/02/1976	VERDADERO	[CÁPSULA DURA]	[VÍA ORAL]	[75.0]	[CINARIZINA]	[FUMARATO DE EST]	[MEDICAMENTO SU]	[N07C - PREPARAD]		

Figura 3: Salida del CSV

Para cada medicamento, representado por un registro en las filas del dataset, se guardan los siguientes campos:

- **Registration Number:** es del tipo numérico; hace referencia al número de registro del medicamento.
- **Name:** de tipo texto; representa el nombre del medicamento.
- **Company:** de tipo texto; contiene el nombre del laboratorio donde se produce el medicamento.
- **Authorization Date:** es del tipo fecha. Hace referencia a la fecha de autorización de la comercialización del medicamento.
- **Commercialized:** de tipo texto; toma el valor *Verdadero* si se ha comercializado y *Falso* si no.
- **Pharmaceutical Dose Form:** es una lista de valores de tipo texto; contiene todas las formas posibles en las que se presenta el medicamento.
- **Routes Administration:** es una lista de valores de tipo texto; representa las vías de administración posibles del medicamento.
- **Strength:** lista de valores de tipo decimal; contiene todas las dosis posibles en las que se vende el medicamento.
- **Active Ingredients:** lista de valores de tipo texto. Presenta los principios activos del medicamento.
- **Excipients:** lista de valores de tipo texto. Contiene todos los excipientes² del medicamento.
- **Characteristics:** lista de valores de tipo texto. Contiene otras características del medicamento como la obligación de una prescripción médica para poder obtenerlo o si afecta de alguna manera a la conducción, etc.

²sustancia inactiva usada para incorporar el principio activo

- ATC Codes: lista de valores de tipo texto. Contiene los códigos ATC de los principios activos del medicamento.
- Images URL: lista de valores de tipo texto; contiene las URL a las imágenes presentadas en la página web.

6. Agradecimientos

Todos los datos guardados en el conjunto de datos obtenido en el proyecto que en este documento se presenta han sido extraídos de la página web del Centro de Información online de Medicamentos Autorizados (Cima).

Para dicha extracción, teniendo en cuenta que el fin del proyecto es meramente educativo, se han cumplido los términos legales teniendo en consideración los aspectos comentados en el *Derechos de propiedad intelectual y de propiedad industrial* del Aviso Legal³ de la misma página, en la que se comentan los siguientes puntos:

- Se autoriza la reproducción total o parcial de los contenidos de la web, siempre que se cite expresamente su origen.
- El usuario queda obligado a mencionar la fecha de la última actualización de los documentos objeto de la reutilización.

De esta manera se presenta el origen de los datos en este mismo apartado, y la Fecha de última Actualización en el README del proyecto.

Por otro lado, y aunque este proyecto no se ha basado en ningún proyecto anterior, se ha podido ver otro tipo de análisis que parecen perseguir un objetivo similar al que aquí se presenta. En primer lugar, hay empresas que prestan un servicio para poder realizar un análisis evolutivo de precios del mercado de medicamentos (Lis Data Solutions). Además, la misma página CIMA presenta un servicio a usuarios para poder descargar el total de la información de la Base de Datos en formato Excel.

7. Inspiración

Como se comentaba en el apartado 1, el objetivo de este proyecto es poder ofrecer a diversos tipos de usuarios una manera de obtener un volumen de información elevado en un conjunto de datos, obteniendo esta información a partir de una fuente fiable, de calidad y actualizada.

Dependiendo del usuario al que se pretende dar servicio, las preguntas que podría responder este conjunto de datos son diferentes:

³<https://www.aemps.gob.es/avisoLegal/#derechos>

Farmacéuticos

En caso de que un farmacéutico atienda a una persona con necesidades detalladas y no tenga el medicamento concreto que se necesita podría preguntarse: ¿qué otro medicamento tiene especificaciones similares y podría servir a esta persona?

Para este tipo de usuario también sería útil poder descargarse los medicamentos por Fecha de Autorización o por si están comercializados o no.

Público general

Un ciudadano que no pertenezca al profesional sanitario puede encontrar útil este conjunto de datos para consultar características de múltiples medicamentos o buscar alternativas a un medicamento en concreto. Por ejemplo, podría responder a la pregunta: ¿qué medicamentos sin receta son similares a otro que la necesita?

Empresas

También podría resultar útil para empresas de packaging. En este caso podrían utilizar el conjunto de datos para analizar a la competencia o establecer su línea de acción. ¿Cómo es el packaging para cada producto en concreto? ¿Existen tendencias en el packaging en función del principio activo, etc?

Data Scientist

Otro posible usuario final es el equipo de científicos o analistas de datos de empresas del sector para poder realizar análisis comparativo entre laboratorios, analizar tendencias de diferentes laboratorios, precios de medicamentos a lo largo del tiempo, etc.

Investigación (industria farmacéutica)

De manera similar al perfil de empresa, el resultado de este proyecto podría también resultar útil para realizar un estudio de la competencia en la industria farmacéutica (laboratorios, centros de investigación), analizando los productos según vía de administración o características, y en función del análisis establecer el propio plan de acción.

Como se comentaba en el apartado 6, existen empresas e iniciativas con objetivos similares al que se tiene en este documento, pero una ventaja clara que presenta este proyecto es la personalización y a la vez generalidad de uso del software. Es decir, el conjunto de datos obtenido mediante el código de este trabajo podría servir a cualquiera de los perfiles

enumerados anteriormente, pero para cada uno de ellos se obtendría información diferente. Por tanto, este proyecto está orientado a que el usuario pueda realizar una búsqueda personalizada.

Por poner un ejemplo, según lo comentado en el apartado 6, CIMA también permite generar un excel con los campos, pero tiene una presentación menos apropiada para tratar los datos y hay campos como *Características*, *Código ATC*, o el enlace a las imágenes que no se obtienen.

8. Licencia

Para este proyecto se ha elegido la licencia de *Creative Commons CC BY-NC-SA 4.0 License*. Esta licencia es idónea por lo siguiente:

- Se debe proveer el nombre del creador del conjunto de datos generado, indicando los cambios que se han realizado.
- Las modificaciones o contribuciones realizadas deben distribuirse bajo los mismos términos.
- No se permite un uso comercial.

Según los puntos anteriores, se establece el reconocimiento del trabajo del autor original y se evita la redistribución del software con fines comerciales.

9. Código

El código del proyecto, así como los resultados, se puede encontrar en el siguiente repositorio:

<https://github.com/rociogm96/medicine-Scraper>

10. Dataset

El CSV se puede encontrar en la web de Zenodo según el siguiente link:

<https://zenodo.org/record/6437472#YlMu2MhBxD8>

Para el conjunto de datos anterior, el DOI es el **10.5281/zenodo.6437472**, y el enlace.

11. Contribuciones

Todos los integrantes del grupo han participado en los diferentes apartados.

CONTRIBUCIONES	FIRMA
Investigación Previa	Juan Luis González Rodríguez Rocío González Martínez
Redacción de las respuestas	Juan Luis González Rodríguez Rocío González Martínez
Desarrollo del código	Juan Luis González Rodríguez Rocío González Martínez