Tipología y Ciclo de Vida de los datos
Práctica 1

Edison Marcelo Muzo Oyana,
aztleclan@gmail.com, emuzo@uoc.edu
Jorge de León,
jorgemariodlc@gmail.com, jdlc@uoc.edu

15 de abril de 2019

1. Contexto

Este conjunto de datos se centra en las enfermedades raras con nomenclatura Orphanet. La enfermedad rara puede ser un trastorno, un síndrome de malformación, un síndrome clínico, una anomalía morfológica o biológica o una situación clínica particular que afecta a menos de 5 personas en una población de 10,000. Aproximadamente existen entre 6,000 y 8,000 enfermedades raras conocidas [8].

Orphanet se encarga de mantener la nomenclatura de las enfermedades raras (ER), esencial para mejorar la visibilidad e impulsar investigaciones que descubran mejores diagnósticos, tratamientos y cuidados para los afectados. Cada enfermedad rara con nomenclatura Orphanet tiene un identificador único y estable, conocido como el número ORPHA.

El portal de Orphanet www.orpha.net permite obtener a los usuarios acceder a la información de las ERs además de los genes, signos y síntomas, discapacidades asociadas a las ERs a través de un formulario de búsqueda.

2. Título para el dataset

Para este trabajo el nombre del dataset será: "**Enfermedades raras**". El nombre es el que mejore define el contenido del dataset.

3. Descripción del dataset

Para realizar este trabajo se aplica la técnica de web scraping en la página de búsqueda de Enfermedades Raras (ER) de Orphanet [9] para recuperar la información asociada a cada ER. El software utilizado en este trabajo es: Python 3.7 [10] y Scrapy [12]. Scrapy es un Framework de código abierto y colaborativo para extraer datos de sitios web. Para instalar Scrapy abrimos una ventana de línea de comandos y ejecutamos el siguiente comando.

En primer lugar, para el rastreador creado con Scrapy respetan el archivo robots.txt se configura la siguiente opción en el fichero *setting.py*:

```
ROBOTSTXT_OBEY = True
```

Luego, cada vez que el rastreador intente descargar una página desde una URL no permitida, verá un mensaje como este:

```
DEBUG: Forbidden by robots.txt: &lt;GET http://website.com/login&gt;
```

Con esta configuración de un rastreador cortes procedemos a recopilar la información de enfermedades raras desde Orphanet. Así, en la página de "**búsqueda de ERs**" accedemos a cada uno de los enlaces de la lista alfabética (*ALPHABETICAL LIST*) de ERs. Cada enlace mostrará una página con un listado de ER. Seguimos cada enlace para acceder a la información particular de cada ER. En la Figura 1 se puede ver los enlaces que se siguen en la página de búsqueda de ERs.

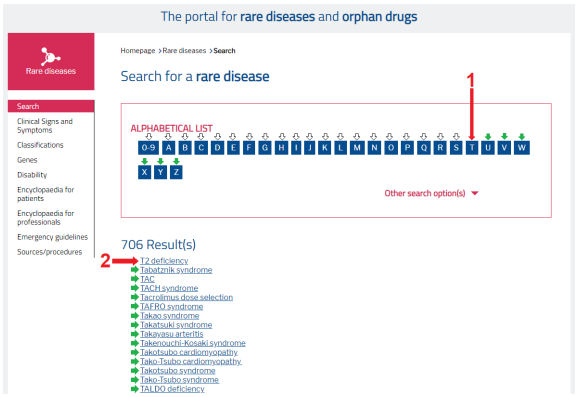

Figura 1: Información Principal enfermedad rara "Toxic oil syndrome".

En cada página de una ER se recupera la siguiente información:

1. Se recupera el nombre de la enfermedad del título de página.
2. Se recupera la información de la ER (OrphanNum, Synonyms, Prevalence, Age Of Onset, ICD10 [3], OMIM [7], UMLS [14], MeSH [5], GARD [1] y MedDRA [4]) de la sección "ORPHA:OrphaNum". Donde OrphaNum es el número orphan de un ER.
3. Se recupera el enlace a PubMed del enlace "Publications in PubMed" en la sección "Additional information".
4. Se recupera el enlace de la página actual de la ER.

Los acrónimos ICD10 [3], OMIM [7], UMLS [14], MeSH [5], GARD [1] y MedDRA [4] hacen referencia a bases de datos, tesauros biomédicos. En las Figuras 2 y 3 se muestra la información que se recupera de cada página de una ER y la Tabla 1 de descripción de los campos.

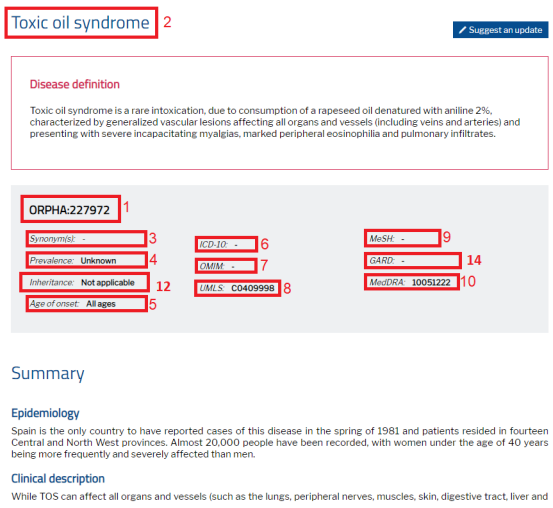

Figura 2: Información Principal enfermedad rara "Toxic oil syndrome".

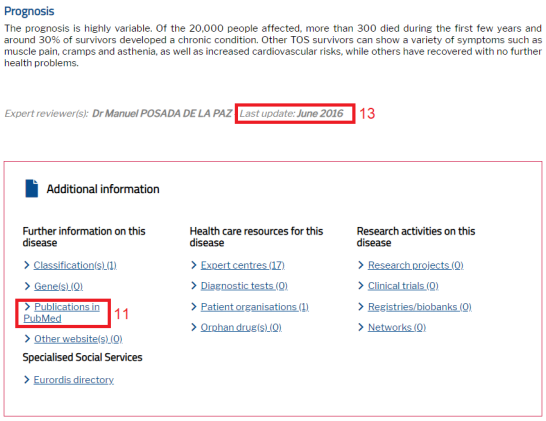

Figura 3: Información Adicional de enfermedad rara "Toxic oil syndrome".

4. Representación gráfica

En primer lugar, para crear un proyecto con Scrapy en la línea de comandos ejecutamos lo siguiente:

```
scrapy startproject orphanetCrawler
```

Renombramos el directorio "orphanetCrawler" a "orphanetWebScraping". Posteriormente para crear el spider en la línea de comandos ejecutamos lo siguiente:

```
cd orphanetWebScraping
scrapy genspider -t crawl orphanet www.orpha.net
```

Como resultado obtenemos la siguiente estructura de directorios y ficheros:

```
orphanetCrawler
        scrapy.cfg
        orphanetCrawler
                __init__.py
                exporters.py      # Exportador a CSV
                items.py
                middlewares.py
                pipelines.py
                settings.py       # Configuración
                spiders           # Directorio de rastreadores
                        orphanet.py  # Rastereador
```

Los ficheros importantes para la construcción del dataset son los siguientes:

1. En el fichero *setting.py* configuramos el comportamiento cortés del rastreador con la siguiente opción:

   ROBOTSTXT_OBEY = True

2. En el fichero *items.py* se establecen los campos que va a tener el dataset.

3. En el fichero *exporters.py* esta el código que exporta la información recolectada a un dataset en formato CSV.

4. En el fichero *orphanet.py* está el código del rastreador que recuperará la información de las ERs desde del portal de Orphanet.

Para crear el dataset ejecutamos en la línea de comandos lo siguiente:

```
scrapy crawl orphanet -o report.csv -t csv -s LOG_FILE=orphanCrawl.log
```

En la Figura 4 se muestra el proceso global de web scraping de este trabajo y en laFigura 5 se muestra la imagen que representa el dataset construido.

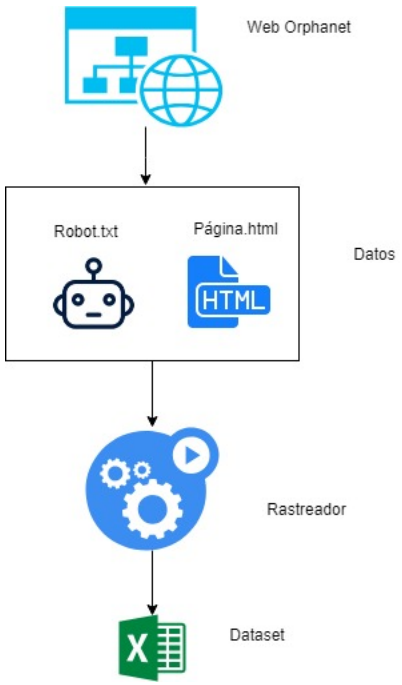

Figura 4: Proceso de web scraping

Figura 5: Enfermedades raras (www.rarediseasereview.org)

5. Contenido

El conjunto de datos consta de quince atributos y 9346 enfermedades obtenidos desde las página web de cada una de las ER, el período de tiempo de los datos dependerá de las actualizaciones que se realicen específicamente

en cada una y del periodo de ejecución del crawler para extraer la información del portal www.orphanet.com.

No	Atributo	Descripción
1	orphan	Número orphan para identificar unívocamente la ER.
2	name	Nombre de la ER.
3	synonyms	Lista de sinónimos separados por coma.
4	prevalence	Frecuencia de aparición.
5	ageOfOnset	Edad de inicio (Neonatal, Infancy, Childhood, Adolescent, Adult y All Ages)
6	ICD10	Identificador de Codificación de diagnósticos y procedimientos.
7	OMIM	Identificador de Conjunto de reglas básicas sobre herencia.
8	UMLS	Identificador de Sistema de lenguaje médico unificado.
9	MeSH	Identificador de Palabras claves de términos médicos.
10	MedDRA	Identificador de Diccionario Médico para actividades reguladoras.
11	PubMed	Enlace a artículos científicos sobre la ER en PubMed.
12	inheritance	Herencia
13	last_updated	Última fecha de actualización de la entrada
14	GARD	Identificador de Centro de información sobre enfermedades genéticas y raras.
15	url	URL correspondiente a la enfermedad

Cuadro 1: Descripción de campos del dataset.

El dataset generado es un fichero en formato CSV donde las columnas están *separadas por un punto y coma*; y el valor de cada columna está *encerrado por comillas dobles*. Los campos únicos campos obligatorios son *orphan*, *name*, y *url*. El resto de los campos son opcionales ya que dependen si la página recolectado dispone de algún valor. Cuando los valores de los campos no vienen informados se representan con el *caracter guión* "-" esto sucede principalmente con aquellos que se obtienen de la información principal (Ver Figura 2).

El dataset actual no ha pasado por un proceso de limpieza de datos. Por tanto, se identifican las siguientes tareas para el proceso de limpieza de datos:

1. Extraer el número del campo "orphan" para ello se tendría que eliminar la cadena "ORPHA:".
2. Cambiar el *carácter guion* "-" por "".
3. Eliminar caracteres extraños o caracteres especiales de html. Por ejemplo, " ".

6. Agradecimientos

El propietario de los datos es Orphanet. Agradecemos el compromiso de mantener la nomenclatura para las enfermedades raras y así mejorar la visibilidad de las mismas.

7. Inspiración

Una de las consecuencias de la baja incidencia de las ERs en la población global es la falta de apoyo y la carencia de recursos para impulsar investigaciones que descubran mejores diagnósticos, tratamientos y cuidados para pacientes que padecen este tipo de enfermedades. La recopilación de un dataset para el ámbito de la investigación académica puede ser relevante.

En el presente dataset tiene los siguientes objetivos:

- Recopilar los posibles sinónimos, edad de inicio y frecuencia de aparición de cada enfermedad rara.
- Mostrar la relación entre la nomenclatura de orphanet y otras bases de datos biomédicas como: ICD10 [3], OMIM [7], UMLS [14], MeSH [5], GARD [1] y MedDRA [4].
- Mostrar el enlace a PubMed que contienen los artículos científicos relacionados con una ER en particular.

8. Licencia

Publicado bajo licencia CC BY-NC_SA 4.0 por los siguientes motivos:

- Se debe reconocer y citar la obra del autor, reconociendo el trabajo y las aportaciones originales.
- Solo para fines no comerciales, ya que las enfermedades raras carecen de recursos suficientes para realizar investigación y desarrollo.
- La publicación de obras se hará bajo una licencia idéntica a la licencia que regula la obra original, por lo que se perpetua la filosofía del autor.

Contribuciones	**Firma**
Investigaciones previas	Edison Muzo; Jorge de León
Redacción de las respuestas	Edison Muzo; Jorge de León
Desarrollo código	Edison Muzo; Jorge de León

Cuadro 2: Grupo de Práctica

9. Código y Dataset

El código y el dataset de la práctica esta en el siguiente repositorio de github:

https://github.com/emmuzoo/orphanetWebScraping.git

10. Referencias

[1] GARD. Gard home page. https://rarediseases.info.nih.gov/, 2019.

[2] Github. Tutorial de github. https://guides.github.com/activities/hello-world, 2019.

[3] ICD10. Icd10 home page. https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/index.html?redirect=/ICD9ProviderDiagnosticCodes/08_ICD10.asp, 2019.

[4] MeDRA. Medra home page. https://www.meddra.org/, 2019.

[5] MeSH. Mesh home page. https://www.ncbi.nlm.nih.gov/mesh, 2019.

[6] United States National Library of Medicine. Pubmed search engine. https://www.ncbi.nlm.nih.gov/pubmed/, 2019.

[7] OMIM. Omim home page. https://www.omim.org/, 2019.

[8] Orphanet. Orphanet home page. https://www.orpha.net/consor/cgi-bin/index.php?lng=EN, 2019.

[9] Orphanet. Search rare disease. https://www.orpha.net/consor/cgi-bin/Disease_Search.php?lng=ES&search=Disease_Search_List, 2019.

[10] Python. python home page. https://www.python.org/, 2019.

[11] Robot.txt. The web robots pages. http://www.robotstxt.org/, 2019.

[12] Scrapy. Scrapy home page. https://scrapy.org/, 2019.

[13] Valdir Stumm. How to crawl the web politely with scrapy. https://blog.scrapinghub.com/2016/08/25/how-to-crawl-the-web-politely-with-scrapy, 2019.

[14] UMLS. Umls home page. https://www.nlm.nih.gov/research/umls/, 2019.

11. Bibliografía

[15] Richard Lawson. *Web scraping with Python*. Packt Publishing Ltd, 2015.

[16] Laia Subirats Maté and Mireia Calvo González. Web scraping. *UOC*, page 54.