PRÁCTICA 1
TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Antonio Fernández Serra,
Noviembre de 2021

CONTEXTO

La presente memoria, trata sobre el proyecto elaborado para la primera práctica de la asignatura Tipología y Ciclo de Vida de los datos obligatoria en el programa de Máster en Ciencia de Datos.

Este trabajo consta de un proyecto de *Web Scraping* (ws) sobre el url de pubmed.ncbi.org a fin de obtener un dataset con información sobre un objeto de investigación biológica. Para ello recabaremos información de la base de datos (bd) de referencia en el mundo de las Ciencias biológicas y biomédicas. Este recurso es provisto por el National Centre of Biotechnology Information (NCBI) [1]. Como misión de esta organización está el conocimiento de la biología molecular, que a partir de un lenguaje de 4 letras, despliega toda la complejidad de la vida. Dentro de este abanico de fenómenos se dan una serie de sútiles patrones de efectos biológicos evidentes e incluso cruciales, por lo que el hecho de disponer de bases de datos bibliográficas de calidad, entre otros recursos computacionales se demuestra de importancia crucial. Así pues en Noviembre de 1988, se establece el NCBI como una rama de la *National Library of Medicine* (NLM) en los *National Institutes of Health* (NIH). Esta combinación aportó al proyecto el conocimiento de la NLM en la construcción y mantenimiento de bases de datos bibliográficas y el NIH la implementación de un laboratorio de biología molecular computacional [2].

En la actualidad este recurso constituye el estándar y referente en la búsqueda bibliográfica en investigación biomédica, constituyendo una herramienta básica para cualquier investigador o investigadora en ciencias de la salud. Tanto es así, que el único recurso que puede hacerle sombra en este ámbito es

Google Académico [3], y principalmente porque tiene una integración mejor con software gestor de referencias bibliográficas como Endnote [4].

A día de hoy, se estima que pubmed contiene más de 33 millones de referencias a artículos científicos, resúmenes de ponencias en congresos y simposios científicos o médicos y libros científicos provenientes de MEDLINE la bbdd de la NLM, cuyo rasgo distintivo es el uso de los Medical Sujects Headings (MeSH), un diccionario jerárquico producido y controlado por la NLM [5] que proporciona una fuente de indexación útil y robusta para las referencias. Una búsqueda bibliográfica exhaustiva, eficiente y bien dirigida es clave en muchas fases de la investigación científica, por lo que todo este ecosistema será ubicuo en el día a día de los científicos y científicas, lo que da idea de su importancia capital en este ámbito.

Una vez fijada la importancia del recurso, se hace evidente que el ws puede ser de mucho ayuda en el uso de pubmed en la potenciación y aceleración de la búsqueda, selección, filtrado y gestión en general de las fuentes bibliográfica en un proyecto de investigación biomédica. Así pues, en el presente trabajo se desarrolla como caso de estudio perfectamente generalizable, la construcción de un *dataset* con información del gen de fusión *TMPRSS2:ERG*, principalmente en oncología, aunque como veremos luego también marginalmente en otros ámbitos. El término de búsqueda está incluido dentro del script, por lo que con cambiarlo en el código podemos hacer el ws sobre otro tema.

Un gen de fusión es un fenómeno biológico que se da cuando existe un proceso patológico (o al menos anómalo), que cambia la estructura cromosómica del DNA nuclear, lo que produce un cambio conformacional, que hace genes que deberían estar muy distantes, se yuxtapongan, unan parte de sus pautas de lectura y formen un gen aberrante, llamado quimérico con parte de un gen y del otro. Si el proceso no es deletéreo forma un nuevo gen que se perpetúa en distintas iteraciones del ciclo celular acabando formando un clon del tumor.

En el caso concreto que estudiamos el gen de fusión *TMPRSS2-ERG* es específico de tumores prostáticos. El descubrimiento de este gen en 2005 constituyó un pequeño terremoto en el campo de la Biología Molecular del cáncer. Por un lado, hasta esa fecha este tipo de genes habían sido extensamente estudiados y caracterizados en tumores sanguíneos y sarcomas. Tanto es así que la translocación *BCR-ABL* uno de los

primeros genes de fusión descritos y que constituye tanto un biomarcador muy específico como una diana terapéutica en ciertos tipos de leucemia [6]. Sin embargo en tumores sólidos nunca se habían descrito este tipo de alteraciones. Cuando en 2005 un grupo de la Universidad de Michigan describió el gen *TMPRSS2:ERG* en aproximadamente la mitad de los casos de un tumor tan frecuente como el cáncer de próstata [7]. Esta referencia bibliográfica puede encontrarse en el dataset adjunto a la práctica. Esta es una de las razones que hacen idónea esta búsqueda, que está muy acotada en el tiempo. Como características, remarcar que este gen va a estar principalmente asociado con Cáncer de Próstata, aunque en los últimos tiempos también se ha estudiado profusamente el primer partner de la anomalía génica (*TMPRSS2)* en el ámbito de la microbiología como el receptor al que se une el virus Cov-19 para entrar en la célula. Otra característica que hace idóneo este término de búsqueda es que tiene una gran cantidad de referencias asociadas, lo que lo hace un buen caso de estudio. Por último que sigue siendo objeto de investigación activa, de hecho durante este año 2021 se han publicado 38 artículos científicos sobre este gen (Tabla 1)

Search query: tmprss2:erg	
Year	Count
2021	38
2020	69
2019	66
2018	71
2017	79
2016	99
2015	101
2014	97
2013	109
2012	62
2011	67
2010	57
2009	55
2008	36
2007	30
2006	13
2005	1
1997	1

Tabla 1: frecuencia absoluta de publicación de artículos científicos sobre el gen de fusión *TMPRSS2:ERG* [8]

Por todas estas razones considero que esta búsqueda en esta página web constituye un buen sustrato para esta práctica centrada en WS, en las coordenadas de interés, dificultad técnica y formato del dataset obtenido.

TÍTULO

Implementación de un dataset de referencias bibliográficas biomédicas: caso de estudio del gen de fusión *TMPRSS2:ERG*.

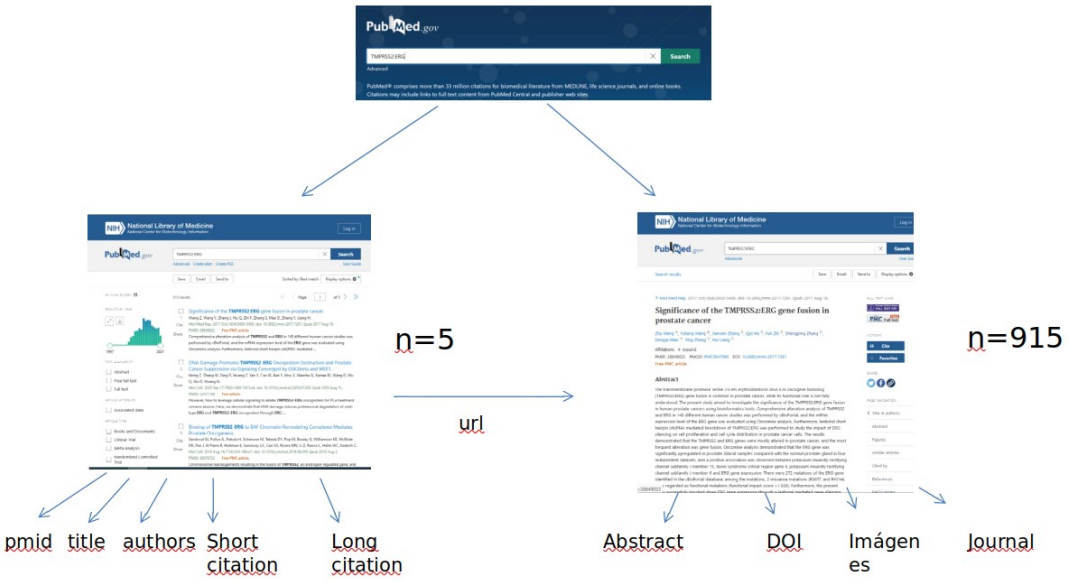
DESCRIPCIÓN DEL DATASET

El dataset consiste en un archivo excel con 915 filas que corresponden a referencias bibliográficas y 9 columnas con la siguiente información de cada una de ellas:

- Pubmed id: número de identificación en la bd pubmed.
- Title: título del trabajo.
- Authors: lista de todos los autores y autoras implicados.
- Long citation: cita del artículo con la que se puede acceder.
- Short citation: cita corta.
- Url: dirección de acceso a cada uno de los artículos.
- Abstract: resumen del trabajo.
- DOI: referencia del número DOI
- Journal: título de la revista científica donde se ha publicado el trabajo.

Aunque el caso de estudio nos lleva a un dataset de referencias sobre el gen *TMPRSS2:ERG*, es generalizable a cualquier término de búsqueda relacionada con ciencias biológicas o biomédicas con lo que obtendríamos un dataset con esta información pero referida a otro objeto de investigación.

REPRESENTACIÓN GRÁFICA



CONTENIDO

El dataset está constituido con un archivo excel compuesto por: el título, autores, id propio de pubmed que sirve para indexar la navegación, cita larga y corta, abstract o resumen de cada uno de los artículos científicos correspondientes al término de búsqueda, DOI y título de la revista científica.

Aunque el caso de estudio se centra en el gen de fusión *TMPRSS2:ERG*, el uso del script es generalizable para el ws de las referencias bibliográficas de cualquier otro término de búsqueda en investigación biomédica o biológica.

Los usos del script son evidentes, al agregar toda la información, pueden filtrarse los artículos de mayor interés por el contenido de su abstract, incluido en el dataset, y navegar directamente con el campo que contiene la url en las diferentes referencias.

El script también contiene la funcionalidad de descarga de imágenes, con el código utilizado podrían descargarse automáticamente todas las imágenes, eventualmente en formato gif, pero con un pequeño cambio podría adaptarse para cualquier otro formato gráfico como png, jpg, etc. Debido a que en esta búsqueda hay más de 5000 imágenes, se descargarán la de un artículo como ejemplo.

Al introducir el término de búsqueda la url nos devuelve una página con las 10 primeras referencias, adaptando la petición pueden obtenerse hasta 200 referencias por página (&size=200) [9]. Aún así nos encontramos con que tenemos que navegar por 5 páginas para completar todas nuestras referencias, la solución a este problema ha sido acceder al código fuente de cada una de las 5 páginas para crear una lista de listas de objetos beautifulsoup sobre los que iteraremos para obtener todos los parámetros. Una solución similar se adoptó en la navegación por los urls específicos de cada uno de los artículos para obtener el abstract, DOI y nombre de la revista en el que se itera y almacena el código fuente de cada página en una lista de listas de objetos bs4, para luego iterar sobre ellos.

AGRADECIMIENTOS

En primer y más evidente lugar el agradecimiento tanto a los NIH como a la NLM y al NCBI, que de modo altruista, han puesto a punto y mantienen todas las infraestructuras para disponer de la información bibliográfica bien indexada para toda la comunidad científica. Este tipo de recursos son fundamentales para el avance de la ciencia dentro de un modelo colaborativo.

En segundo lugar a la misma comunidad científica que produce todo el conocimiento contenido en estas referencias a base de unos esfuerzos tanto en conseguir los fondos para realizar las investigaciones y llevarla s a cabo como preparar los manuscritos, someterse a un proceso de revisión por pares y sufragar los gastos de publicación. En el *dataset* incluido han contribuido varios miles de investigadores e investigadoras que han llevado el descubrimiento del gen de fusión desde sus inicios en ciencia básica hasta su aplicación como diana terapéutica en cáncer de pulmón, aumentando drásticamente la esperanza de vida de los pacientes [10].

La justificación de este proyecto es la automatización y optimización de la exploración de la bibliografía existente en las fases de planificación de nuevos proyectos científicos, o bien en la discusión de los resultados obtenidos.

En cuanto al respeto de los principios éticos y legales se ha actuado en el marco de la legislación y buenas prácticas en ws.

En primer lugar no se ha navegado sobre los sites especificados en el archivo robots.txt [11] (Anexo 1). Por otro lado en el script se incluyen una serie de cabeceras para evitar ser baneados cuando la web detecte que se trata de un bot [12]:

```
user_agent_list = [
'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_5) AppleWebKit/605.1.15 (KHTML, like Gecko)
Version/13.1.1 Safari/605.1.15',
'Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:77.0) Gecko/20100101 Firefox/77.0',
'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_5) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/83.0.4103.97 Safari/537.36',
'Mozilla/5.0 (Macintosh; Intel Mac OS X 10.15; rv:77.0) Gecko/20100101 Firefox/77.0',
'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/83.0.4103.97 Safari/537.36']
```

También se ha implementado en el script un tiempo de espera para no saturar el servidor:

driver.implicitly_wait(1)

Este comando sirve para espaciar las peticiones http. Esto es debido a que los servidores web están adaptados para la navegación humana, en la que el proceso de navegar entre los links tarda segundos, mientras que un *web scraper* envía todas las peticiones simultáneamente.

En este caso, el mismo diseño del proyecto hace que no sea necesario implementar defensas contra las trampas de araña, porque el dominio de la búsqueda está perfectamente delimitado como puede apreciarse en el apartado de representación gráfica, la petición se hace en la misma url de pubmed, que nos devuelve todos los resultados coincidentes con el criterio de búsqueda fragmentados en lotes de 200. Cada una de estas referencias contiene un link a una página específica en la que se obtiene el abstract.

En este caso no ha sido preciso aceptar ningún contrato explícita o implícitamente vía login en el que se restrinjan en modo alguno las actividades de ws. Por tanto no existe incumplimiento de condiciones contractuales.

Los derechos de autor en artículos científicos están por una serie de reglas que dependen de si el trabajo ha sido publicado en una revista científica o es un pre-print. En el primer caso el tipo de licencia depende de cada revista [13]. En cualquier caso los datos contenidos en este dataset son de libre acceso y no entran dentro del ámbito del cuerpo del paper, con el que si que estaríamos entrando en terreno protegido.

En el script se incluye el login a la bd pubmed con mi usuario.

En cualquier caso se han seguido las directrices de verificar las condiciones de uso, rastrear sólo información pública, no causar daño (sobrecargando el servidor) y usar la información de manera justa.

INSPIRACIÓN

El presente trabajo parte de una necesidad real en el ámbito de la investigación biomédica. La búsqueda crítica y síntesis de material bibliográfico es un paso ineludible en el curso de cualquier proyecto de investigación. Mientras que los procesos de evaluación crítica y síntesis no son automatizables, e incluso constituyen parte del transfondo más filosófico de la investigación, no es así con los procesos de búsqueda y recolección de referencias que es un proceso tedioso, repetitivo y automatizable. El presente proyecto de ws persigue facilitar al máximo esta parte de la búsqueda bibliográfica, teniendo entre sus objetivos proporcionar un conjunto exhaustivo y no sesgado de referencias bibliográficas.

LICENCIA

La licencia Creative Commons Zero (CC0) es ningún derecho reservado, todo lo que se encuentra bajo esta licencia puede ser usado, modificado, explotado incluso con interés comercial sin ni siquiera necesidad de citar la fuente [14].

Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0): la única diferencia entre esta cláusula y la anterior es la cláusula Share-alike. Bajo la licencia CC0, cualquiera que adapte un trabajo y lo redistribuya, puede hacerlo con cualquier licencia, mientras que en este caso se restringe que esa redistribución siga el mismo modelo de licencia [15]. Además esta licencia restringe el uso comercial del trabajo derivativo de este dataset.

Attribution-ShareAlike 4.0 International (CC BY-SA 4.0): A diferencia de la anterior esta si que permite el uso comercial de trabajos derivados [16].

Open Database License: permite compartir, crear y adaptar libremente las bases de datos. Todos los procesos derivados de ellas deben conservar estos atributos y con una licencia que garantice lo mismo [17].

CC0 1.0 Universal: supone la renuncia a cualesquiera derechos cubiertos por el copyright en la medida en que lo permita la ley [18]

CÓDIGO

https://github.com/afernandezse/Practica_1_M2.851-_Tipologia_y_ciclo_de_vida

DATASET

DOI: 10.5281/zenodo.5648435

En la url https://zenodo.org introducir la referencia 5648435 en la caja de búsqueda

BIBLIOGRAFÍA

[1] https://www.ncbi.nlm.nih.gov/

[2] https://www.ncbi.nlm.nih.gov/home/about/mission/

[3] https://scholar.google.es/schhp?hl=es

[4] https://endnote.com/

[5] https://www.nlm.nih.gov/mesh/meshhome.html

[6] https://en.wikipedia.org/wiki/Philadelphia_chromosome

[7] https://pubmed.ncbi.nlm.nih.gov/16254181/

[8] https://pubmed.ncbi.nlm.nih.gov/?term=tmprss2%3Aerg&sort=date

[9] https://pubmed.ncbi.nlm.nih.gov/?term=tmprss2%3Aerg&sort=date&size=200

[10] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8554320/pdf/ott-14-5161.pdf

[11] https://pubmed.ncbi.nlm.nih.gov/robots.txt

[12] https://www.scrapehero.com/how-to-fake-and-rotate-user-agents-using-python-3/

[13] https://biblioguias.uam.es/produccion/repositorio/derechos_autor

[14] https://oshl.umh.es/2016/03/15/que-significa-cc0/

[15] https://meta.wikimedia.org/wiki/Open_Content_-_A_Practical_Guide_to_Using_Creative_Commons_Licences/The_Creative_Commons_licencing_scheme

[16] https://creativecommons.org/licenses/by-sa/4.0/

[17] https://opendatacommons.org/licenses/odbl/summary/

[18] https://creativecommons.org/publicdomain/zero/1.0/

[19] https://creativecommons.org/licenses/by/4.0/

ANEXO 1

Archivo robots.txt

User-agent: *

```
Crawl-delay: 1
Disallow: /api
Disallow: /rss
Disallow: /advanced/adv-suggestions/
Disallow: /terms/
Disallow: /addToHistory/
Disallow: /deleteHistory/
Disallow: /downloadHistory/
Disallow: /deleteHistoryRecord/
Disallow: /historyCacheExists/
Disallow: /*/references/
Disallow: /*/citations/
Disallow: /*/export/
Disallow: /*/citedby/
Disallow: /*/similar/
Disallow: /*/adj-nav/
Disallow: /ajax/
Disallow: /clipboard/
Disallow: /clipboard-next-page/
Disallow: /health/
Disallow: /deep-health-abstract/
Disallow: /deep-health-search/
Disallow: /deep-health-auth/
Disallow: /error/400/
Disallow: /error/403/
Disallow: /error/404/
Disallow: /error/500/
Disallow: /results-export-ids/
Disallow: /results-export-search-data/
Disallow: /results-export-email-by-search-data/
Disallow: /results-export-search-by-year/
Disallow: /send-email/
Disallow: /list-existing-collections/
Disallow: /add-to-existing-collection/
Disallow: /create-and-add-to-new-collection/
Disallow: /toggle-favorites-collection/
Disallow: /list-bibliography-delegates/
Disallow: /add-to-bibliography/
Disallow: /create-saved-search/
Disallow: /create-rss-feed-url/
Disallow: /searches/
Disallow: /collections/
Disallow: /more/
Disallow: /suggestions/
Disallow: /try-search-term/
```

```
Disallow: /rss-feed/

Sitemap: https://pubmed.ncbi.nlm.nih.gov/sitemap?p=index.xml
```

ANEXO 2

print(whois.whois('https://pubmed.ncbi.nlm.nih.gov'))

{
"domain_name": "NIH.GOV",
"registrar": null,
"whois_server": null,
"referral_url": null,
"updated_date": null,
"creation_date": null,
"expiration_date": null,
"name_servers": null,
"status": "ACTIVE",
"emails": "ResponsibleDisclosure@hhs.gov",
"dnssec": null,
"name": null,
"org": null,
"address": null,
"city": null,
"state": null,
"zipcode": null,
"country": null
}