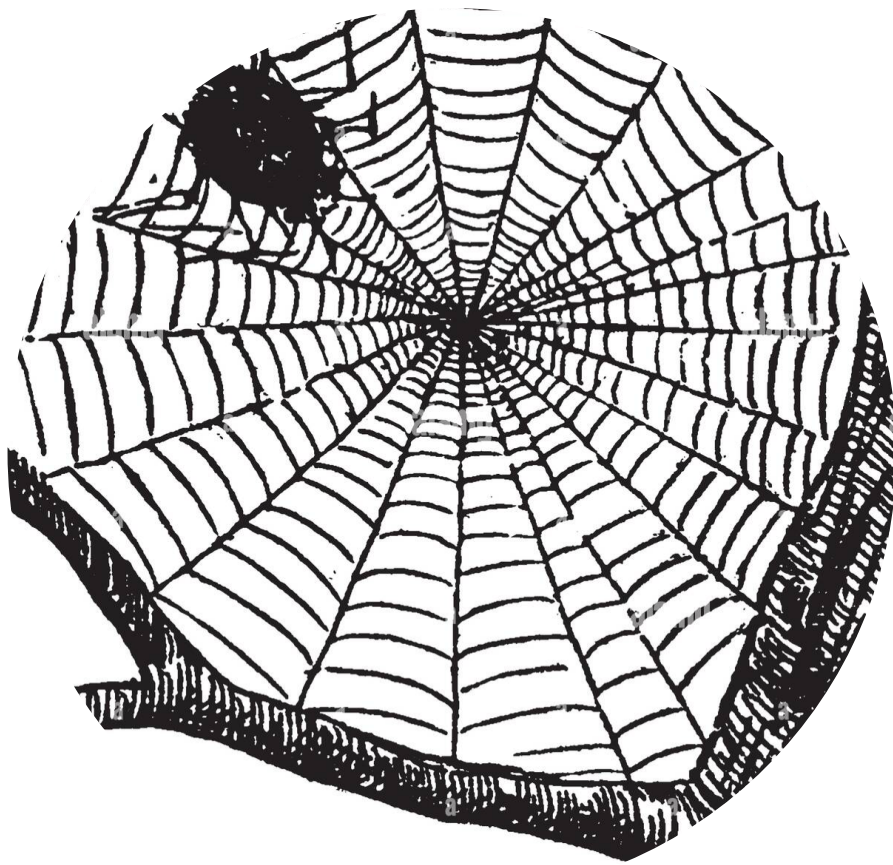


Web scraping
Catálogo de la Editorial Anagrama



Germán del Cacho Salvador
Enrique Callejas Castro

Marzo de 2022

Máster Universitario en Ciencia de Datos
Universitat Oberta de Catalunya

Índice

1. Contexto	2
2. Título del dataset	2
3. Descripción del dataset	3
4. Representación gráfica	3
5. Contenido	6
6. Agradecimientos	7
7. Inspiración	7
8. Licencia	9
9. Código	9
10. Dataset	10
11. Recursos	10
12. Contribuciones	10

1. Contexto

El mundo editorial ha experimentado importantes cambios en los últimos años. El uso masivo de dispositivos electrónicos como tabletas y libros electrónicos ha hecho que el libro en papel haya disminuido sus ventas, dando paso a otros formatos como el *ebook* o el audiolibro. A través de este trabajo queremos realizar un estudio de caso de la editorial Anagrama. Anagrama es una editorial de referencia a nivel nacional. Fue fundada en 1969 y, desde entonces, ha publicado 4000 títulos aproximadamente. Hemos escogido su propia página web (<https://www.anagrama-ed.es/>) porque consideramos que es la que tiene información más precisa sobre sus propios títulos. Además, en el contexto de la práctica, representa un reto ya que los datos no pueden ser extraídos mediante una API. Por ello, tenemos que usar técnicas de *web scraping*. Asimismo, el diseño de la página nos permite aprender técnicas avanzadas como la navegación a través de *webs* paginadas. De forma complementaria, hemos enriquecido el dataset obtenido de Anagrama con la información que proporciona Amazon España respecto de las ventas de cada libro¹. Esta información podría ser utilizada para optimizar las ventas de la editorial. En este sentido, podrían utilizarse técnicas de procesamiento de lenguaje natural y de aprendizaje automático para conocer si aspectos tales como el género del libro (que podría extraerse de su sinopsis) o la elección de la portada (mediante el análisis de las portadas recopiladas) tienen incidencia en las ventas. Como en el caso de Anagrama, tampoco Amazon dispone de una API oficial que permita obtener fácilmente esta información², por lo que hemos tenido que utilizar nuevamente técnicas de *web scraping*, en este caso mediante *Selenium*, ya que los productos de Amazon solo son accesibles desde formularios web que es preciso rellenar.

2. Título del dataset

El dataset tiene por título *coleccion_anagrama*, que ilustra por sí mismo el contenido.

- *coleccion*: ilustra que el dataset recoge toda la información del catálogo de títulos.
- *anagrama*: hace referencia a la procedencia de los registros.

Además, dado que Anagrama es una editorial conocida, consideramos que no es necesario incluir la palabra *libros* en el título. De este modo, mantenemos un título más sintético y fácil de manejar en el código.

¹ Para cada producto, Amazon proporciona el lugar que ocupa dicho producto en distintos rankings, según su categoría (ranking de libros más vendidos, de libros en español más vendidos, de libros de suspense más vendidos...).

² Amazon solo facilita el acceso a su API oficial, Product Advertising API 5.0, a propietarios de páginas webs que se comprometen a promocionar sus productos.

3. Descripción del dataset

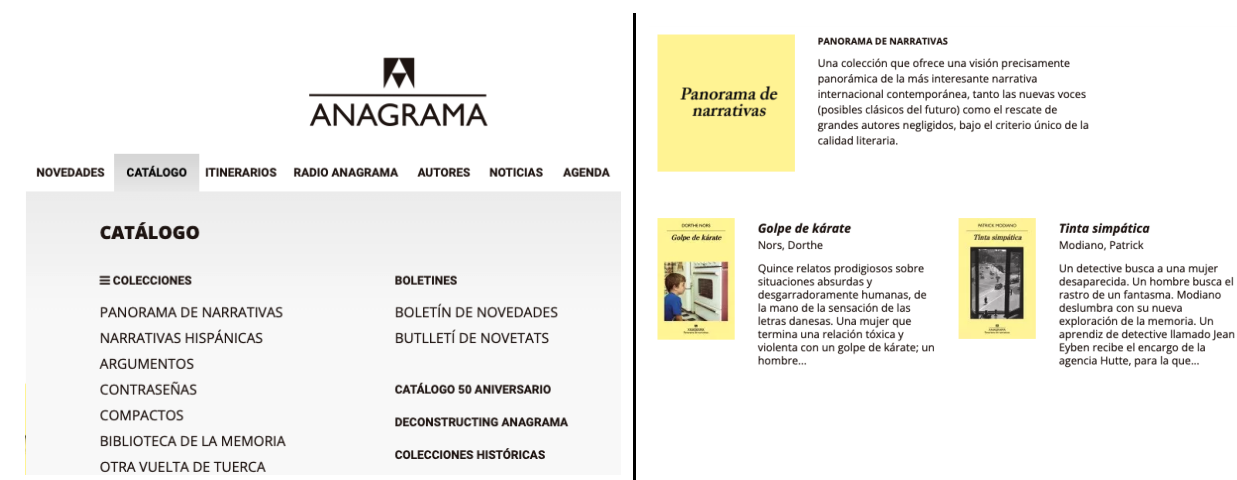
Este trabajo presenta la recogida de datos del catálogo online de la Editorial Anagrama, enriquecido con los datos de ventas de Amazon. En concreto, contiene un total de 3819 títulos de 1063 autores diferentes recogidos en 20 colecciones. El catálogo recoge información general de los libros, algunas de sus variables son la autoría, resumen, número de páginas o el precio.

4. Representación gráfica

A continuación, se presenta un resumen visual del proyecto y del *dataset*.

La página web de Anagrama ofrece un catálogo de sus libros ordenados por colecciones (Figura 1, izq.) y, dentro de cada una de ellas, la relación de todos sus títulos (Figura 1, der.). Dado la gran cantidad de títulos de cada colección, estos se encuentran distribuidos en distintas páginas en las que se puede navegar a través de botones tipo *siguiente/anterior*.

Figura 1. Captura de pantalla de la página web de Anagrama.



Fuentes:

- (1) <https://www.anagrama-ed.es/colecciones>
- (2) <https://www.anagrama-ed.es/coleccion/panorama-de-narrativas>

Para poder acceder a la información completa de cada libro, es necesario clicar en el título deseado, que nos dirige a su página concreta. En la Figura 2 podemos ver un ejemplo.

Figura 2. Ejemplo de página web del título *Golpe de kárate*



PAPEL E-BOOK

ISBN	978-84-339-8117-2
EAN	9788433981172
PVP CON IVA	17,9 €
NÚM. DE PÁGINAS	144
COLECCIÓN	Panorama de narrativas
CÓDIGO	PN 1077
TRADUCCIÓN	Victoria Alonso
PUBLICACIÓN	27/04/2022

Golpe de kárate

Dorte Nors

Quince relatos prodigiosos sobre situaciones absurdas y desgarradoramente humanas, de la mano de la sensación de las letras danesas.

Una mujer que termina una relación tóxica y violenta con un golpe de kárate; un hombre que, mientras su esposa duerme, indaga en internet acerca de la historia de una asesina psicópata; la criada mexicana de una sofisticada pareja danesa instalada en Manhattan que debe bregar con un tomate gigante; un niño que, cargado de buenas intenciones, acaba horneando a un pato vivo; una persona con discapacidad que acepta un engaño en su búsqueda de bondad humana; un hijo que descubre perplejo la fragilidad de su admirado padre; un hombre que decide aplicar el budismo a sus relaciones laborales con sorprendentes consecuencias... Estos son algunos de los singulares personajes que protagonizan los quince relatos aquí reunidos.

Cuentos tan breves como contundentes, en ocasiones perturbadores y en otras perturbadoramente hilarantes, que escrutan con demoledora agudeza comportamientos humanos. Historias que nos hablan de soledades, anhelos, angustias, perplejidades, fragilidades, desconciertos y otras muchas realidades cotidianas.

Con una economía de medios deslumbrante, un preciso control del ritmo y una endiablada capacidad de observación, la autora nos propone una jugosa galería de peculiares personajes, un portentoso mapa de situaciones absurdas y desgarradoramente humanas.



Dorte Nors

Dorte Nors (Herning, 1970) es licenciada en Literatura e Historia del Arte por la Universidad de Aarhus, y una de las voces más originales y aplaudidas de la literatura danesa actual. Es autora de cuatro novelas y de un volumen de relatos, *Golpe de kárate*, con el que dio el salto internacional. Ha publicado textos en revistas como *Harper's* y *Boston Review*, y en 2013 un cuento suyo fue el primero de un escritor danés en el *New Yorker*. En 2014 recibió el Premio Per Olov Enquist. En Anagrama ha aparecido la novela *Espejo, hombre, intermitente*: «Llena de miniaturas vitales contadas con ironía y profundidad, Dorte Nors cuenta la soledad urbana. Lo cómico de la soledad» (Rosa Belmonte, ABC); «Nors es una miniaturista deliciosa que se apoya en la ficción experimental para retratarnos la vida interior y el aislamiento de los seres anónimos de mediana edad» (Ángeles López, *La Razón*).

Fuente: https://www.anagrama-ed.es/libro/panorama-de-narrativas/golpe-de-karate/9788433981172/PN_1077

Por tanto, el *dataset* recoge los principales datos de la página individual de cada libro.

En cuanto a los datos de Amazon, las páginas de los productos son accesibles desde distintos formularios web. En concreto, Amazon dispone de un formulario para la búsqueda avanzada de libros, que permite introducir el ISBN o el EAN:

Figura 3. Formulario web para libros de Amazon

Advanced Search

Books Search

Keywords

Author

Title

ISBN(s)

Publisher

Subject All Subjects

Condition All Conditions

Format All Formats

Reader Age All Ages

Language All Languages

Pub. Date All Dates Month Year

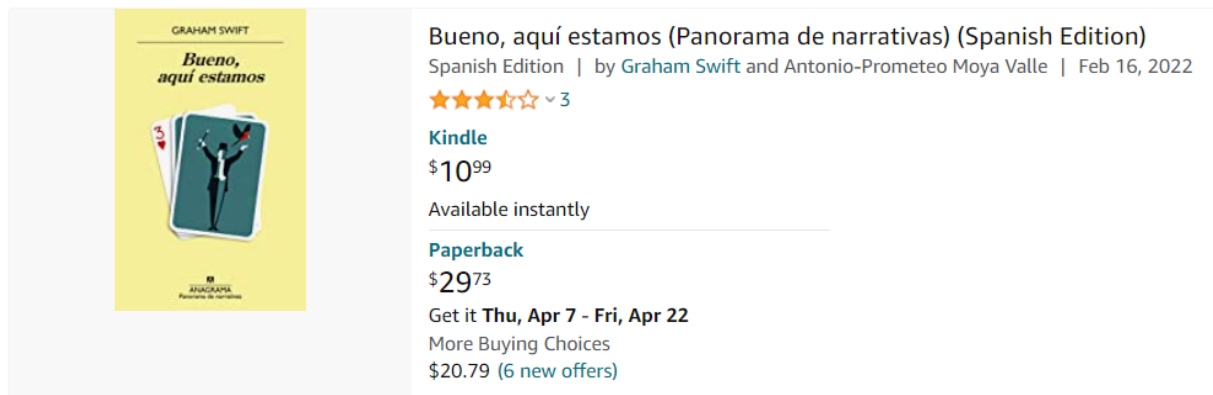
Sort Results by: Featured

Search

Fuente: <https://www.amazon.com/advanced-search/books>

Mediante estas claves, que proporciona Anagrama en su catálogo, se accede a una ventana que recoge los resultados de la búsqueda:

Figura 4. Resultados de búsqueda en Amazon



Fuente: https://www.amazon.com/s?i=stripbooks&rh=p_66%3A9788433981127&s=relevanceexprank&Adv-Srch-Books-Submit.x=5&Adv-Srch-Books-Submit.y=6&unfiltered=1&ref=sr_adv_b

Clicando en el producto, es posible acceder a la página que contiene los datos del libro, entre los que se encuentra la información comercial que nos interesa recopilar:

Figura 5. Página del producto (información comercial)

Best Sellers Rank: #1,142,923 in Kindle Store (See Top 100 in Kindle Store)
 #4,430 in Literature & Fiction in Spanish
 #13,686 in Spanish Language Fiction
 #76,614 in Libros en español

Fuente: https://www.amazon.es/En-casa-los-sue%C3%B1os-narrativas/dp/8433980904/ref=sr_1_4?__mk_es_ES=%C3%85M%C3%85C5%BD%C3%95%C3%91&crd=1K2K8T3RB3XUQ&keywords=editorial+anagrama&qid=1648030497&srefix=editorial+anagrama%2Caps%2C102&sr=8-4

Los distintos tops de ventas para cada libro, según categoría, han sido almacenados en una lista. En función del análisis a realizar, sería necesario añadir tantas columnas como tops registrados, incluyendo como valores el lugar que ocupa cada libro en cada top (nulo en el caso de que el top se corresponda con una categoría diferente). En la medida en que este análisis no se va a realizar, ya que excedería el alcance de esta práctica y sería necesario disponer de un equipo completo de científicos de datos, el *dataset* no ha sido preprocesado en el sentido indicado. El interés de la información, aun sin utilizarse, justifica su inclusión. En la siguiente figura (Figura 6) se puede ver la cabecera del *dataset* final, que consta de catorce columnas.

Figura 6. Cabecera del *dataset* final

	isbn	ean	código	título	autor	precio	páginas	fecha_publicación	colección	traducción	sinopsis	ebook	tops_amazon
0	978-84-339-8106-6	203513482	PN 1066	En verano	Knausgård, Karl Ove	18.9	400	2021-11-24	Panorama De Narrativas	Asunción Lorenzo, Kirsti Baggethun	La portentosa culminación del Cuarteto de las ...	True	['3,322,967 en Tienda Kindle', '6,137 en Biogr...
1	978-84-339-8113-4	203513550	PN 1073	Historia de los abuelos que no tuve	Jablonka, Ivan	21.9	424	2022-02-23	Panorama De Narrativas	Agustina Blanco	La historia de los abuelos del autor, muertos ...	False	['1,835,602 en Libros', '63,596 en Memorias (L...
2	978-84-339-8107-3	203513489	PN 1067	Desde dentro	Amis, Martin	24.9	624	2021-11-17	Panorama De Narrativas	Jesús Zulaika Golcochea	Un libro ambicioso y deslumbrante que mezcla v...	True	['804,658 en Tienda Kindle', '641 en Biografía...
3	978-84-339-8110-3	203513519	PN 1070	Sed	Nothomb, Amélie	17.9	128	2022-02-02	Panorama De Narrativas	Sergi Pàmies	Una apasionante y nothombiana reelaboración de...	True	['377,300 en Tienda Kindle', '844 en Literatur...
4	978-84-339-8111-0	203513526	PN 1071	Nacido de ninguna mujer	Bouysse, Franck	19.9	304	2022-02-02	Panorama De Narrativas	Rosa Alapont Calderaro	Un manuscrito. Una joven enfrentada a un desti...	True	['2,090,463 en Libros', '99,130 en Libros en e...

5. Contenido

El *dataset* cuenta con unas dimensiones de 3819 filas y 14 columnas, incluyendo el género (que ha sido derivado del nombre de los autores mediante una librería específica). Sus campos son (*formato en cursiva*):

- **isbn**: código normalizado internacional para libros (International Standard Book Number) (*string*)
- **ean**: es el código ISBN sin los guiones intermedios (*integer*)
- **código**: código del título (*string*)
- **título**: nombre de la obra (*string*)
- **autor**: autor/a de la obra (*string*)
- **precio**: precio de venta al público en euros (*float*)
- **páginas**: número de páginas de la obra (*integer*)
- **fecha_de_publicación**: fecha en la que se publicó la edición en venta (*datetime*)
- **colección**: colección de la editorial a la que pertenece (*string*)
- **traducción**: nombre del traductor/a
- **sinopsis**: resumen de la obra (*string*)
- **ebook**: libro disponible en formato electrónico (*boolean*)
- **género del autor/a**³: hombre/mujer (*string*)
- **tops_amazon**: posición del libro en cada ranking de Amazon (*list*)

Procedimiento y herramientas

El proceso de recogida de datos se llevó a cabo el día 16 de marzo de 2022. Se realizó mediante *web scraping*. En concreto se emplearon librerías de Python como BeautifulSoup y Selenium. La recogida se estructuró en dos fases:

³ Variable creada a posteriori a partir del nombre del autor/a mediante la librería gender-guesser

Scraping de los títulos: se realizó un primer *scraping* que tenía por objetivo recoger los enlaces individuales de todos los títulos. Para ello, primero obtuvimos los enlaces de todas las colecciones haciendo un *scraping* en <https://www.anagrama-ed.es/colecciones> y, después, realizamos un *scraping* dentro del enlace de cada colección en el que extrajimos los enlaces referentes a los libros. Una vez recogidos los enlaces de todos los títulos del catálogo, se realizó el *scraping* de manera individual de cada uno de ellos, accediendo a su enlace correspondiente. Entre una petición y la siguiente, se estableció un descanso proporcional al tiempo de respuesta para no saturar el servidor. Cada petición recogía la información relativa al título de la obra, autoría, precio, etc. y la cargaba en un diccionario de Python, que a su vez se incluía en una lista con el resto de los libros. Cabe destacar que algunos enlaces ofrecían la información incompleta. Por ejemplo, algunos de ellos no tenían el número de páginas o la traducción, por lo que imputamos “NA” (*Not Available*) en los campos no disponibles. También recopilamos las portadas, que almacenamos en un directorio específico.

Scraping de los datos de ventas de Amazon: Una vez obtenido el catálogo de Anagrama, extrajimos los códigos “ean”, que cumplen en nuestra base de datos funciones de clave primaria. A partir de cada “ean” accedimos a la página de Amazon España para cada libro y recopilamos la información de ventas mediante Selenium. Empleamos dicha librería ya que no es posible acceder a los productos de Amazon mediante la manipulación de urls que cumplen un determinado patrón, sino que es preciso rellenar un formulario web e interactuar con el navegador.

Buenas prácticas y ética

El proceso se realizó de forma ética. Algunas de las medidas que adoptamos fueron: (1) espaciar las consultas para no saturar el servidor *web*; (2) revisar los aspectos legales de la página web; (3) rastrear únicamente la información pública. Anagrama no dispone de un archivo robots.txt, pero sí Amazon. Las directrices de dicho archivo, que se encuentra en la carpeta del proyecto, fueron respetadas.

6. Agradecimientos

Agradecemos a la Editorial Anagrama ofrecer sus datos en abierto, así como a Amazon España.

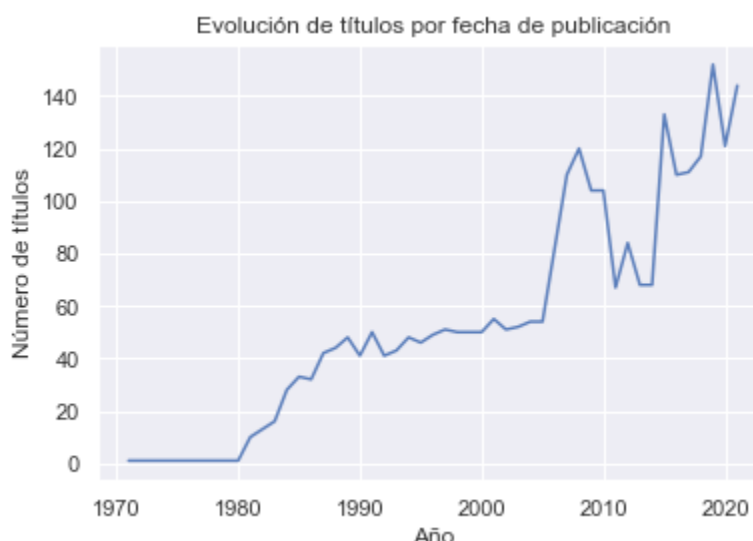
7. Inspiración

Este trabajo representa un estudio de caso de la evolución del ámbito editorial en los últimos años. Creemos que el análisis del *dataset* generado puede ayudar a responder a las siguientes preguntas:

¿Cómo ha evolucionado el número de títulos publicados anualmente?

Podemos observar que el número de títulos evoluciona de forma constante, excepto en 2008 aproximadamente, que experimenta una caída que se prolonga hasta 2014. Creemos que esto se puede deber a la gran crisis económica de 2008, que también tuvo un impacto muy negativo en el ámbito cultural. Además, en torno a este periodo, irrumpen nuevos dispositivos personales (e.g., ordenadores portátiles, *smartphones*) que suplen la función de entretenimiento de los libros tradicionales en formato de papel. Es posible que a partir de 2014 se recupere la tendencia de crecimiento por la incorporación de los libros en formato electrónico (Figura 4).

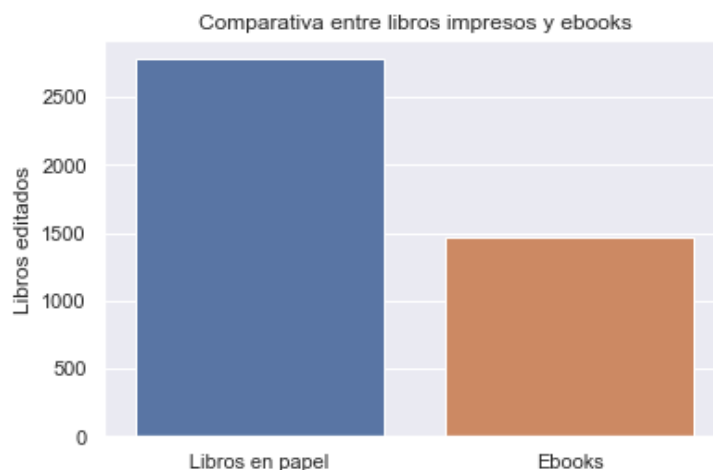
Figura 4. Evolución del número de títulos por año



¿Qué proporción de libros se encuentran disponibles en formato electrónico?

Observamos que ya encontramos una gran proporción de libros en formato electrónico, aunque todavía se encuentra muchos que podrían ser comercializados en formato digital.

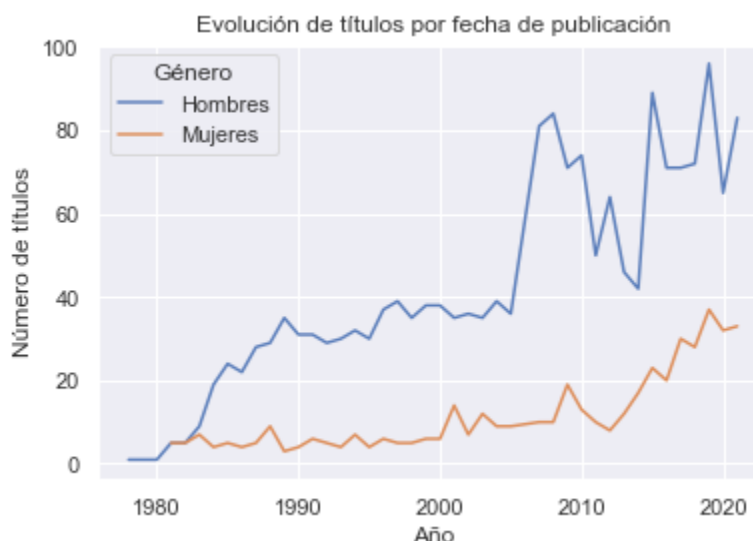
Figura 5. Frecuencia de títulos disponibles en formato electrónico.



¿Cuál es la proporción y la evolución de los títulos escritos por mujeres y por hombres a lo largo del tiempo?

Observamos que el número de títulos de hombres es mucho mayor que el de mujeres, aunque en los últimos años parece que ambas proporciones comienzan a acercarse. En el caso de la tendencia temporal de los hombres observamos picos de subida y de bajada más acusados, mientras que la tendencia de mujeres parece más estable (Figura 6).

Figura 6. Evolución de títulos publicados por mujeres y por hombres



8. Licencia

El trabajo se presenta bajo licencia **Attribution 4.0 International** (CC BY 4.0), lo que implica que se puede **compartir** en cualquier medio o formato así como **adaptar** para cualquier medio, incluso comercial. Hemos escogido esta licencia porque es la que permite que el trabajo de disemine en mayor medida y que, por tanto, pueda ayudar o inspirar a más personas. Respecto a la **atribución**, debe darse el crédito apropiado, proporcionando un enlace a la licencia e indicando los cambios realizados, en caso de que los hubiera.

9. Código

El código para obtención del dataset *anagrama_coleccion* puede encontrarse en los archivos: *scraper.ipynb*. Además, se proporciona el código de los análisis en *coleccion_anagrama_analysis.ipynb*.

10. Dataset

El *dataset* puede encontrarse en los formatos .db y .csv en la carpeta *data*. Además, se ofrece en formato .csv y se encuentra disponible en la plataforma [zonodo.org](https://zenodo.org/doi/10.5281/zenodo.6379996) con el siguiente DOI asociado: 10.5281/zenodo.6379996.

11. Recursos

Calvo, M., Pérez, D., Subirats, L. (2019). *Introducción al ciclo de vida de los datos*. UOC
Lawson, R. (2015). *Web scraping with Python*. Packt Publishing Ltd.

12. Contribuciones

CONTRIBUCIONES	FIRMAS
Investigación previa	
Redacción de las respuestas	
Desarrollo del código	