

¿Cómo podemos capturar los datos de la web?

Práctica 1

William Gabriel Granda Betancourt

Oscar Augusto Diaz Triana

UOC

Universitat Oberta
de Catalunya

Índice

- Descripción de la práctica
- Desarrollo de la práctica
 - Contexto
 - Título
 - Descripción del dataset
 - Representación Gráfica
 - Contenido
 - Propietario
 - Inspiración
 - Licencia
 - Código
 - Dataset
 - Video
- Recursos
- Criterios de valoración
- Referencias



Descripción Práctica 1

Presentación

En esta práctica se elaborará un caso práctico orientado a identificar y extraer datos relevantes para un proyecto analítico, empleando herramientas específicas de web scraping. El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en un sitio web

Competencias

En esta PEC se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para resolverlo.
- Capacidad para aplicar las técnicas específicas de web scraping.

Objetivos

Los objetivos concretos de la práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes cuyo tratamiento aporta valor a una empresa y la identificación de nuevos proyectos analíticos.
- Saber identificar los datos relevantes para llevar a cabo un proyecto analítico.
- Capturar datos de diferentes fuentes de datos (tales como redes sociales, web de datos, o repositorios).
- Actuar según los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar la capacidad de búsqueda.

Desarrollo de la Práctica

Contexto

En el contexto de esta práctica, se busca recolectar datos relevantes sobre las ventas de laptops y computadores en España, con el objetivo de utilizar esta información en un proyecto analítico posterior. El sitio web elegido para la extracción de datos es Amazon España (<https://www.amazon.es/>), debido a que es uno de los principales sitios de comercio electrónico de este país, con una amplia variedad de productos y una gran cantidad de datos disponibles públicamente.

La información obtenida a través del web scraping puede ser de gran utilidad para analizar las tendencias de ventas en distintos productos, identificar los productos más populares, evaluar la satisfacción de los clientes con ciertos productos, entre otros posibles análisis. Además, esta información puede ser de interés para empresas que venden productos en Amazon España, ya que les puede permitir identificar oportunidades de mejora en su oferta de productos, evaluar la competencia, entre otros aspectos.

La recolección de datos a través del web scraping en el sitio web de Amazon España se justifica por la gran cantidad de datos relevantes que se pueden obtener y la utilidad que estos pueden tener para futuros análisis y proyectos analíticos, por ejemplo, en el análisis de ventas de productos, para obtener información valiosa y apoyar en la toma de decisiones en estrategias de marketing.

Título.

Ventas de laptops y computadores en Amazon España durante abril del 2023.

Descripción del dataset

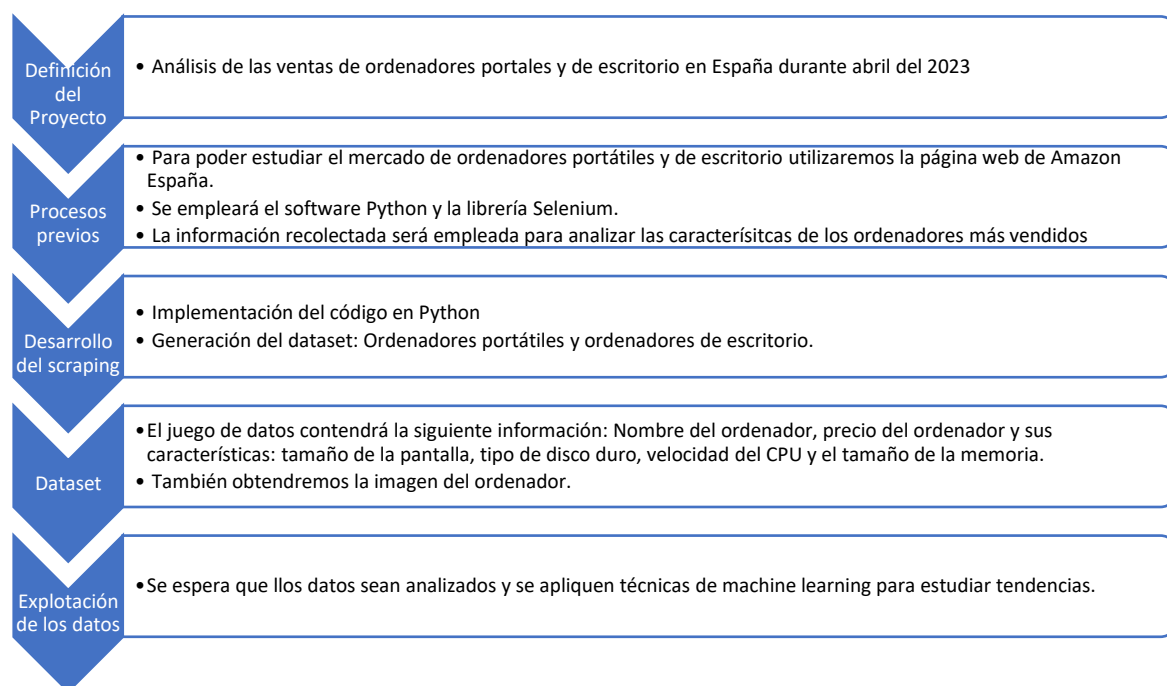
El dataset extraído de la página de Amazon España se titula "*Ventas de laptops y computadores en Amazon España durante abril del 2023*", son datos de la categoría de electrónica y la subcategoría informática. En particular, aquellos relacionados con laptops y computadores publicados durante el mes de abril de 2023. Los datos incluyen nombre del computador, el precio, tamaño de pantalla, descripción del disco duro, velocidad del cpu y tamaño de la memoria, la valoración promedio de los clientes y la imagen de cada computador.

Con el dataset extraído de Amazon España sobre ventas de laptops y computadores durante el mes de abril de 2023, se pueden realizar diversas análisis. Por ejemplo, se pueden identificar los productos más vendidos en esta categoría durante ese periodo y analizar las características que los hacen más populares entre los clientes. También se podría investigar la relación entre la valoración promedio de los clientes y las ventas de los productos, para determinar si hay una correlación entre la satisfacción del cliente y el éxito de ventas de un producto. Además, se podría realizar un análisis de imagen para ver cómo la apariencia de los productos influye en su popularidad y ventas. Todo esto podría ayudar a los fabricantes y vendedores de laptops y computadores a tomar decisiones informadas

sobre cómo mejorar la calidad, el diseño y la promoción de sus productos para aumentar sus ventas y satisfacer a los clientes. Para extraer estos datos, es necesario realizar web scraping a la página de Amazon España.

Representación gráfica

A continuación se realiza presenta un esquema grafico del dataset y del proyecto elegido.



Contenido

El juego de datos cuenta con lo siguientes campos:

- 1) **Nombre:** Este campo identifica el nombre del ordenador portátil o del ordenador de escritorio. Este dato es válido desde la fecha de scraping hasta que se borre la publicación en Amazon.
- 2) **Precio:** Este campo representa el precio de venta en euros de los ordenadores. Este campo es válido desde la fecha de scraping hasta que se borre la publicación de Amazon o se actualicen los precios.
- 3) **Pantalla:** Este campo indica el tamaño de la pantalla en pulgadas del ordenador. De igual forma, este dato es válido desde la fecha de scraping hasta que se borre la publicación en Amazon.
- 4) **Disco:** Este campo indica la descripción del disco duro, por ejemplo, SSD hace referencia a unidades de estado sólido. También, este dato es válido desde la fecha de scraping hasta que se borre la publicación en Amazon.
- 5) **CPU:** Indica la velocidad del CPU medida en GHz. Válido desde la fecha de scraping hasta que se borre la publicación en Amazon.
- 6) **Memoria:** Indica el tamaño de memoria del ordenador, se mide en GB. Este dato es válido desde la fecha de scraping hasta que se borre la publicación en Amazon.

- 7) **Valoración:** Este campo indica la calificación por estrellas de un producto utilizando modelos de aprendizaje de máquina en lugar de un simple promedio y toma valores entre 0 y 5.
Este dato es válido desde la fecha de scraping hasta que se realice una nueva compra y se recalcule la calificación.
- 8) **Imágenes:** Para cada ordenador que extraemos también descargamos su imagen y la guardamos con el formato **Nombre_Precio** para identificar a que ordenador corresponde. El formato de la imagen es JPG.

También, presentamos los primeros cinco registros del dataset obtenido al aplicar scraping para los ordenadores portátiles:

	Nombre	Precio	PANTALLA	DISCO	CPU	MEMORIA	VALORACION
0	SAMSUNG Galaxy Book3 - Laptop 15,6" FullHD (In...	999,00	15.6	SSD	1	1	3.9
1	Acer Swift SF514-55T - Ordenador Portátil 14" ...	1.169,00	14	SSD	1.7	8	4.4
2	Lenovo Legion 5 Gen 6 - Ordenador Portátil 15....	749,00	15.6	SSD	4.4	1	3.5
3	Lenovo IdeaPad 1 Gen 7 - Ordenador Portátil 15...	1.399,00	15.6	SSD	3.5	8	4.1
4	Acer Nitro 5 AN515-58 - Ordenador Portátil Gam...	349,00	15.6	SSD	[]	1	4.4
5	Acer Chromebook 314-1H - Ordenador Portátil 14...	799,00	15.6	SSD	2.8	8	4.4

Estos datos se extrajeron en a inicios del mes de abril de 2023 .

Propietario

Amazon España es una filial de la compañía estadounidense Amazon, una de las mayores empresas de comercio electrónico a nivel mundial. Amazon España ofrece una amplia variedad de productos a través de su plataforma en línea, incluyendo libros, electrónica, moda, hogar y jardín, entre otros. Además, Amazon España ofrece servicios como Amazon Prime, que permite a los clientes acceder a envíos gratuitos y rápidos, así como a una amplia selección de contenidos de entretenimiento en línea.

Como propietario de los datos, Amazon España ha proporcionado información detallada sobre los productos vendidos en su plataforma, lo que permite realizar un análisis exhaustivo de las ventas y tendencias en el mercado español. Esta información es de gran interés para analistas y empresas que deseen conocer mejor el comportamiento del mercado y las preferencias de los consumidores en España. Amazon España ha seguido los principios éticos y legales en la recolección y gestión de los datos, garantizando la privacidad y protección de la información personal de sus clientes.

Hay diferentes autores que han realizado uso del web scraping para cumplir con los objetivos de su investigación. A continuación, se relacionan algunos:

Los autores Chong, Ch'ng, Liu y Li en el año 2017), realizaron un estudio que se centra en investigar las contribuciones del marketing promocional en línea y las revisiones en línea como predictores de la demanda de productos de consumo en Amazon.com. Para ello, se utilizó una arquitectura de Big Data que incorporó el análisis de redes neuronales para predecir si variables de revisión en línea y variables de marketing promocional en línea influyen en la demanda de productos electrónicos en Amazon. El web scraping se utilizó para recolectar los datos electrónicos de Amazon.com y poder realizar el análisis de redes

neuronales. La importancia de este estudio radica en la comprensión de cómo las revisiones en línea y el marketing promocional en línea pueden influir en las demandas de productos, lo que puede ser útil para los profesionales del marketing. En relación con la práctica de web scraping que se está realizando, este estudio muestra cómo la recolección de datos de un sitio web puede ser utilizada para el análisis de datos y para entender mejor el comportamiento del consumidor en línea.

Este estudio, realizado Etumnu et al., (2020), por tiene como objetivo investigar si las reseñas en línea, las estrategias de promoción en línea y los sentimientos de las reseñas de los usuarios pueden ayudar a predecir las ventas de productos. Para ello, los autores diseñaron una arquitectura de big data y utilizaron agentes Node.js para realizar web scraping en Amazon.com. Luego, los datos obtenidos fueron preprocesados para su análisis sentimental y de redes neuronales. Los resultados mostraron que aunque todas estas variables pueden predecir las ventas de productos, la interacción entre ellas es más importante que las variables individuales. Este estudio es relevante para los practicantes, ya que les permite comprender cómo las reseñas y estrategias de promoción en línea pueden influir en las ventas de productos. Además, la arquitectura de big data y el enfoque de análisis predictivo utilizado pueden ser útiles para futuras investigaciones que busquen predecir las ventas de productos en entornos en línea.

Según, Etumnu, (2022), en un estudio reciente (Gil et al., 2020) se encontró que el envío gratuito puede perjudicar a un minorista en línea si su estrategia óptima es cobrar tarifas de envío. En este sentido, se realizó un análisis para conocer si esta observación se aplica a Amazon. Para ello, se utilizó datos de Amazon Canada y se encontró que los productos con envío gratuito reciben más ventas en promedio. El aumento en las ventas se debe probablemente a que los clientes de Amazon tienen reticencias a pagar gastos de envío, y los productos con envío gratuito tienen precios más bajos y una calificación promedio más alta, lo que impulsa la demanda. Para recopilar los datos necesarios para este análisis, se utilizó web scraping.

En cuanto a los pasos seguidos para actuar de acuerdo con los principios éticos y legales, se han considerado las normas de ética y privacidad de datos, así como la normativa europea de protección de datos personales (RGPD). Se ha asegurado que la información recolectada es de dominio público y no viola ninguna normativa legal o ética. Además, se ha garantizado la confidencialidad de la información de los usuarios de la página web y no se ha utilizado ningún tipo de información personal en el dataset obtenido.

Cabe destacar que, al ser una práctica académica, no se utilizará el conjunto de datos obtenido con fines comerciales ni se compartirá con terceros sin previa autorización del propietario.

Inspiración

Este conjunto de datos extraído de Amazon España sobre las ventas de productos en la categoría de electrónica, en particular, laptops y computadoras durante el mes de abril de 2023, podría ser interesante por varias razones.

En primer lugar, estos datos podrían ser valiosos para los vendedores de laptops y computadoras que deseen obtener información sobre las preferencias de los consumidores y la demanda del mercado. Al analizar los datos, los vendedores podrían determinar qué modelos son los más populares, qué características son más valoradas por los consumidores y cómo se comparan los productos de diferentes marcas.

Además, los datos también podrían ser útiles para los consumidores que buscan comprar una nueva laptop o computadora. Al ver la valoración promedio de los clientes y la imagen de cada producto, los consumidores podrían hacer comparaciones y tomar decisiones informadas sobre qué modelo comprar.

En términos de preguntas que se pueden responder con estos datos, algunos ejemplos podrían ser:

- ¿Cuáles son las marcas y modelos de laptops y computadoras más vendidos durante el mes de abril de 2023 en Amazon España?
- ¿Qué características son las más valoradas por los consumidores en la categoría de laptops y computadoras?
- ¿Hay alguna relación entre la valoración promedio de los clientes y la imagen de cada producto y las ventas de los mismos?

Tabla comparativa entre la práctica y los trabajos presentados en el apartador 6:

Análisis	Enfoque	Objetivo	Fuente de datos	Técnicas utilizadas	Resultados
Análisis de ventas de laptops y computadoras en Amazon España en abril de 2023 <u>Práctica 1</u>	Mercadotecnia y consumo	Determinar las preferencias de los consumidores y la demanda del mercado, y proporcionar información útil tanto para los vendedores como para los consumidores	Datos extraídos de Amazon España	Análisis de datos para determinar las marcas y modelos más vendidos, las características más valoradas por los consumidores, y si hay relación entre la valoración promedio de los clientes y la imagen de cada producto y las ventas de los mismos	Proporciona información útil tanto para los vendedores como para los consumidores
Estudio sobre las revisiones en línea y el marketing promocional en línea como predictores de la demanda de productos	Mercadotecnia	Investigar las contribuciones del marketing promocional en línea y las revisiones en línea como predictores de la	Datos obtenidos a través del web scraping en Amazon.com	Análisis de redes neuronales para predecir si las variables de revisión en línea y variables de marketing promocional en línea	Las revisiones en línea y el marketing promocional en línea pueden influir en las demandas de productos

de consumo en Amazon.com		demanda de productos de consumo		influyen en la demanda de productos electrónicos	
Estudio sobre cómo las reseñas en línea, las estrategias de promoción en línea y los sentimientos de las reseñas de los usuarios pueden ayudar a predecir las ventas de productos	Mercadotecnia y análisis predictivo	Investigar si las reseñas en línea, las estrategias de promoción en línea y los sentimientos de las reseñas de los usuarios pueden ayudar a predecir las ventas de productos	Datos obtenidos a través del web scraping en Amazon.com	Análisis sentimental y de redes neuronales para predecir las ventas de productos	La interacción entre las variables es más importante que las variables individuales para predecir las ventas de productos
Análisis de cómo el envío gratuito afecta a las ventas de productos en Amazon	Mercadotecnia	Investigar si el envío gratuito afecta a las ventas de productos en Amazon	Datos obtenidos a través del web scraping en Amazon Canada	Análisis de los productos con envío gratuito y su relación con las ventas, comparando los precios y las calificaciones	Los productos con envío gratuito tienen precios más bajos y una calificación promedio más alta, lo que impulsa la demanda

Licencia

Después de analizar las opciones de licencia, consideramos que la licencia más adecuada para el dataset resultante sería la licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0).

Esta licencia permite que el dataset sea compartido, adaptado y utilizado por cualquier persona, siempre y cuando se atribuya adecuadamente la fuente original y se utilice únicamente con fines no comerciales. Además, esta licencia permite que cualquier adaptación o trabajo derivado creado a partir del dataset también sea compartido bajo la misma licencia, lo que garantiza que cualquier nueva creación también pueda ser utilizada sin fines comerciales y se atribuya adecuadamente a la fuente original.

La elección de esta licencia se debe a que el dataset contiene información valiosa que puede ser utilizada por estudiantes, investigadores y cualquier persona interesada en el análisis de datos en el ámbito de las computadoras. Al mismo tiempo, esta licencia garantiza que el dataset no sea utilizado con fines comerciales, lo que protege los derechos de autor y la propiedad intelectual de los datos originales.

Código

La implementación del código para obtener el dataset se realizó en Python, empleando las siguientes librerías:

- Selenium - Versión: 4.8.3
- Pandas - Versión: 1.2.4
- Numpy - Versión: 1.20.1
- Módulo **re**, para realizar operaciones con expresiones regulares.

En primer lugar, utilizamos la librería Selenium para poder automatizar el proceso de scraping y utilizamos el siguiente WebDriver: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/111.0.0.0 Safari/537.36

La librería Selenium nos permitió gestionar y aceptar las cookies que se generan al visitar la página de Amazon. También, nos permite gestionar correos y contraseñas, en este proceso se tuvo un inconveniente, en ocasiones al momento de ingresar el usuario en Amazon, nos lleva a distintas páginas, en específico a dos. Por ello, implementamos en el código bloques **try** y **except** para manejar estos errores. Si se presenta la siguiente interfaz:



Entonces se ejecuta el bloque **try** y se envía el correo y contraseñas, mientras que si se presenta la siguiente interfaz:





Entonces, damos click en iniciar sesión y se envía el correo y la contraseña.

Luego, obtenemos todas las categorías de productos en Python y escogemos la categoría de computadoras, primero enviamos la palabra *laptops* para poder buscar los ordenadores portátiles en venta.

En algunos casos, resultaba más fácil emplear la búsqueda por XPATH o por CSS SELECTOR.

Al momento de extraer los precios, se presentó otro inconveniente:

	
<p>HP 14s-fq0004ns - Ordenador portátil de 14" Full HD (AMD Athlon Silver 3050U, 8GB RAM, 512GB SSD, AMD Radeon...)</p> <p>4,2 ★★★★★ (266)</p>	<p>Lenovo IdeaPad 3 CB 14M836 Chromebook Gen 6 - Portátil 14" FullHD (MediaTek MT8183, 4GB RAM, 32GB eMMC, Arm Mali-G...)</p> <p>4,3 ★★★★★ (232)</p>
	<p>249,00€</p> <p>50% de descuento en Antivirus</p>

Ciertos productos no tenían el precio, se estudió la estructura y existen tres publicaciones en la parte superior sobre las que extraemos el nombre del ordenador, pero no su precio. Para solucionar esto empleamos una condición **IF** que permite comparar el tamaño de la lista donde almacenamos los nombres con el tamaño de la lista donde almacenamos los precios y así asignamos correctamente el precio con cada computador.

Al momento de extraer las características del ordenador como velocidad del CPU y tamaño de la pantalla, se presentó otro problema. Al utilizar el XPATH obtenemos toda la información en conjunto, como se puede observar en la siguiente imagen:

```
Tamaño de pantalla
15.6 pulgadas
Descripción del disco duro
SSD
Velocidad CPU
1 GHz
Tamaño de memoria
8 GB
```

Para poder reconocer el valor que le corresponde a cada campo (PANTALLA, DISCO, CPU, MEMORIA) empleamos expresiones regulares.

Finalmente, una vez que identificamos la estructura para la primera página, generalizamos el proceso empleando un lazo **FOR**, también identificamos en la página a través del buscador CSS SELECTOR el botón que nos permite navegar entre páginas. Así, podemos obtener la información de los ordenadores ejecutando únicamente el script de Python e indicando el número de páginas. A lo largo del código que se encuentra dentro del lazo

FOR se puede encontrar funciones que permiten generar tiempos de espera para evitar que la página se sature.

Dataset

Dataset Publicado en formato CSV en Zenodo, se incluyó una breve descripción de este.

<https://doi.org/10.5281/zenodo.7823916>



Vídeo

Se comparte el enlace del driver donde está el video explicativo de la práctica

<https://drive.google.com/drive/folders/1tOUCtET3qzI1-YvnnnEEOJSYRHnuEUIV?usp=sharing>

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

- Subirats, L., Calvo, M. (2018).
- Web Scraping. Editorial UOC.
- Masip, D. (2019). El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
- Tutorial de GitHub <https://guides.github.com/activities/hello-world>.

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los apartados es la siguiente:

Apartado	1	2	3	4	5	6	7	8	9	10	11
Puntos	0.25	0.25	0.25	0.5	1	1.5	1.25	0.5	2	2	0.5

Criterios que se tomarán en cuenta para la valoración de la práctica:

- Idoneidad de las respuestas (deberán ser claras y completas).
- Complejidad del sitio web elegido para la extracción de datos. Es importante tener en cuenta que la complejidad será un factor que se evaluará y dependerá tanto del sitio elegido como del análisis realizado en la práctica.
- Síntesis y claridad, a través del uso de comentarios, del código resultante.
- Presentación adecuada de los datos.
- Organización y claridad de los documentos de entrega final.
- Completitud de los documentos requeridos para la entrega final.
- Seguimiento de recomendaciones para el buen uso del web scrapin

REFERENCIAS

- Chong, A. Y. L., Ch'ng, E., Liu, M. J., & Li, B. (2017). Predicting consumer product demands via Big Data: the roles of online promotional marketing and online reviews. *International Journal of Production Research*, 55(17), 5142–5156. <https://doi.org/10.1080/00207543.2015.1066519>
- Etumnu, C. E. (2022). Free shipping. *Applied Economics Letters*. <https://doi.org/10.1080/13504851.2022.2094871>
- Etumnu, C. E., Foster, K., Widmar, N. O., Lusk, J. L., & Ortega, D. L. (2020). Does the distribution of ratings affect online grocery sales? Evidence from Amazon. *Agribusiness*, 36(4), 501–521. <https://doi.org/10.1002/AGR.21653>
- Subirats, L., & Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. (2019). El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd.
- Simon Munzert, C., Rubba, C., Meißner, P., & Nyhuis, D. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
- GitHub Guides. (n.d.). Hello World. Retrieved March 31, 2023, from <https://guides.github.com/activities/hello-world/>

Contribuciones	Firma Integrantes
Investigación previa	WGGB, OADT
Redacción de las respuestas	WGGB, OADT
Desarrollo del código	WGGB, OADT
Participación en el vídeo	WGGB, OADT