

MineríaLibre 2023

Autor:

- Andrade, Cristian Adrián

Resumen

Este proyecto llamado 'MineríaLibre' facilita la minería de datos en artículos vendidos en el sitio de ventas MercadoLibre en base a cualquier búsqueda que se haga, discriminando por condición de producto; si es nuevo, usado, o ambos. El conjunto de datos en su integridad se adapta en la medida que avanza el 'raspado' para incorporar cada uno de los artículos recabados. El resultado final es una planilla de cálculo llena de aquellos registros.

Introducción

Cuando extraje mi primer set de datos con web scraping para limpiar, programé al algoritmo para que trajera los registros que cumplieran con mis exigencias. De esta forma, rescaté muchos artículos de la eliminación. Pero, que hubiera pasado si tomaba los artículos tal y como venían, con todas sus columnas y sin omisión. Hubiese sido un gran desafío limpiarlo. Y una gran oportunidad para practicar limpieza.

Y qué tal, sumado a lo anterior, personalizar el set usando cualquier búsqueda que el usuario realice.

Entonces surge este proyecto llamado 'MineríaLibre', una herramienta diseñada para extraer conocimiento alojado en la plataforma de comercio electrónico MercadoLibre. MineríaLibre es una herramienta que permite la minería de datos precisa y eficiente de productos brindando un puntapié inicial a aquellos usuarios amantes de los datos.

Definición de objetivos

Objetivo General

Desarrollar en Python una aplicación que logre recabar todos los artículos y sus especificaciones técnicas siempre en base a una búsqueda puntual, cometida por el usuario y discriminando por el estado del artículo.

Objetivos específicos

- Investigar sobre cómo están dispuestos los datos en MercadoLibre.
- Diseño de una solución que permita realizar una búsqueda y adaptarse a la condición escogida por el usuario.

MineríaLibre

Este proyecto denominado MineríaLibre es un programa cuyo algoritmo recabará todos los artículos de una búsqueda dada en MercadoLibre y le entregará al usuario un set de datos en formato planilla de cálculo para que haga con él una limpieza, análisis o lo que crea conveniente.

Lo que destacaría de este proyecto es que el algoritmo es capaz de adaptarse a cualquier cantidad de campos de información que tenga el artículo, sean, por ejemplo, cinco en uno o veinte en el próximo.

Deberá ser robusto y resistente a errores, sobre todo porque estamos hablando de una aplicación que usa la internet en casi todo su recorrido, y la inestabilidad del lado servidor jamás puede ser subestimada.

Objetivos y alcances de la aplicación

El objetivo del trabajo de esta investigación es desarrollar una aplicación que haga una búsqueda en MercadoLibre, extraiga datos de páginas HTML, los ingrese a un set de datos, y a su vez que este set se adapte a todos los campos presentes en un artículo. Esto último quiere decir que el set debe reformarse conforme avance en los artículos. Esta reforma es altamente probable que suceda debido a que cada vendedor es único y recae en cada uno de ellos aportar más o menos datos de su producto.

El fin último del proyecto será que el usuario genere sus propios sets de datos de la temática que prefiera buscar y la condición del producto que quiera elegir cuando sea posible.

Justificación

Los sets de datos, en la vida real, nunca vienen limpios de base. Si fuese así, la letra 'T' de los procesos ETL o ELT no existiría. Usar los sets generados por este programa es una gran oportunidad para aquellos que deseen practicar la limpieza de datos bajo la temática que deseen y para artículos de retail.

El web scraping es indiscutiblemente más rápido que visitar cada página, copiar y pegar manualmente cada dato en una planilla o un CSV. Una cosa para recordar es que 'MineríaLibre' corrige sus columnas para hacer entrar todos los registros de manera automática, ahorrándole al usuario aún más tiempo.

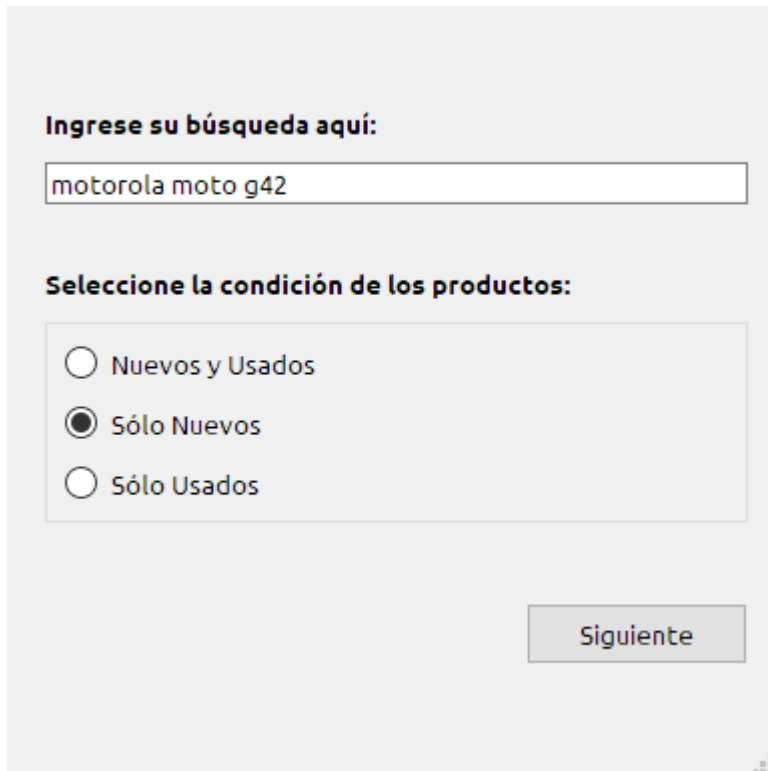
Finalmente, si alguien quiere comparar todo con el todo, con grandes cantidades de registros al mismo tiempo y en unos minutos en vez de horas de navegación, esto es para ellos. Pero si desea hacer pequeñas comparaciones con pocos artículos, este programa, con honestidad, no es conveniente.

Funcionalidades

En esta sección hablaremos de las funcionalidades del sistema con el fin de conocer las funciones soporta, los datos asociados, las restricciones impuestas y cualquier otro factor que pueda influir en su construcción.

Versión con interfaz gráfica

Todo inicia con una búsqueda que escribirá el usuario. Debe ingresar una o varias palabras clave. Además, debe aclarar si quiere sólo nuevos, sólo usados o ambos a la vez:



Ingrese su búsqueda aquí:

Seleccione la condición de los productos:

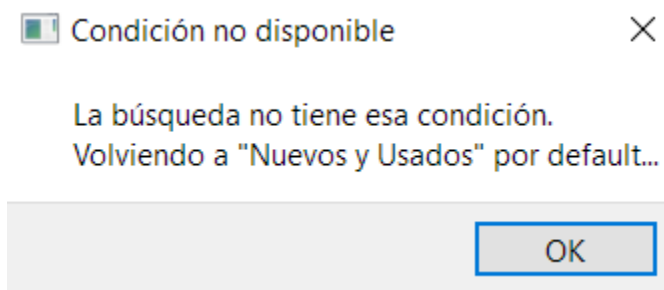
☐ Nuevos y Usados

☒ Sólo Nuevos

☐ Sólo Usados

Siguiete

En caso de que la búsqueda no disponga de nuevos o de usados, se lo hará saber y a cambio no discriminará por condición:



Condición no disponible

La búsqueda no tiene esa condición.
Volviendo a "Nuevos y Usados" por default...

OK


A continuación, le informa de la cantidad de artículos que encontró. En base a este dato, le pregunta al usuario cuántos artículos quiere en su set de datos:

**Hay exactamente 16 artículos con esa condición.
Puede traerlos todos o parte de ellos**

Ingrese 0 si desea todos los artículos ó ingrese una cantidad entre 1 y 16:

¡Comenzar!

Mientras sucede la recolección, le aparecerá este cartel anunciando que aguarde por el fin de la recolección y el tiempo estimado.

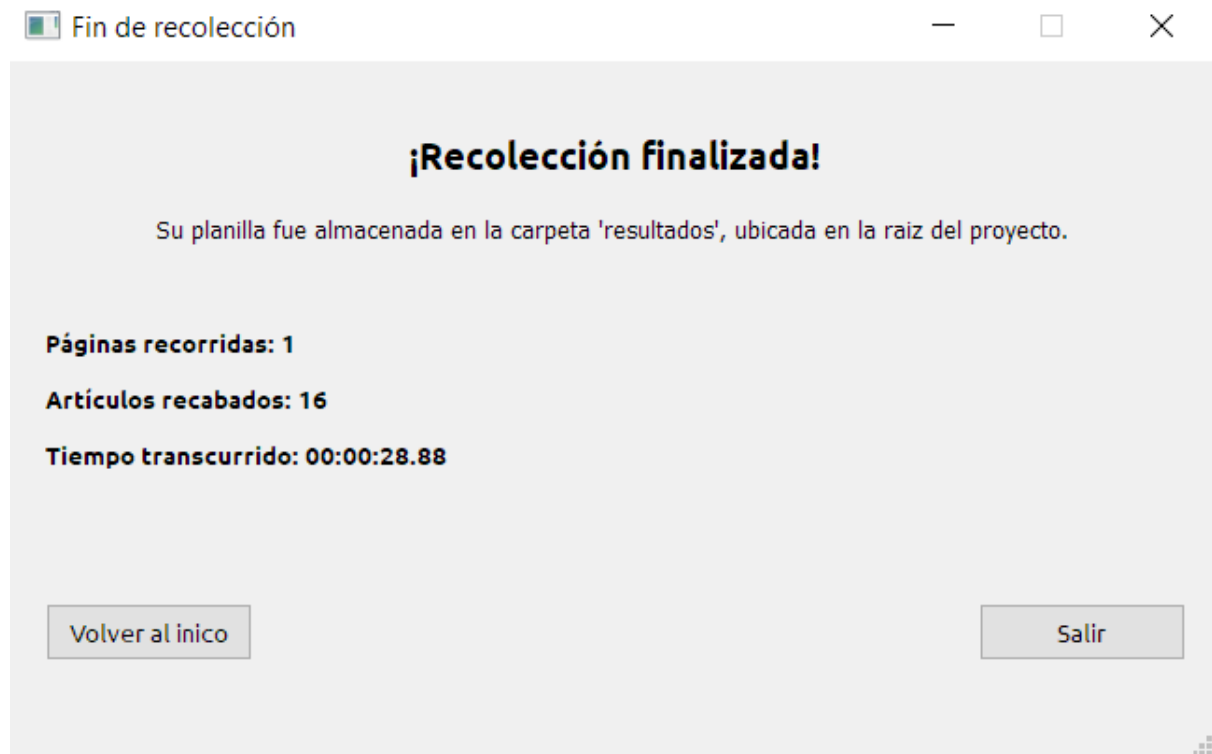
 Espere

— □ ×

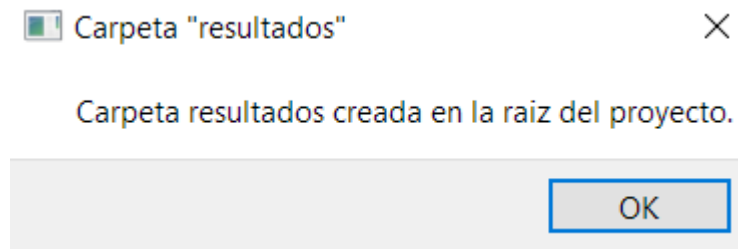
Espere mientras se recolectan los articulos...
Tiempo estimado: 00:00:23

33%


Finalmente, el programa le informa que la planilla con la recolección se depositó en la carpeta 'resultados'. Además, menciona cuántas páginas se visitaron, cuántos artículos se recabaron y el tiempo total de ejecución.



Si la carpeta 'resultados' no fue creada con antelación, se le informará que ya fue creada:



This PC > Documents > Metodología de la investigación > MineríaLibre > resultados

Name	Date modified	Type	Size
 motorola moto g42 - Sólo Nuevos	11/10/2023 09:57	Microsoft Excel W...	15 KB

Puede volver a realizar una recolección si lo desea, llevándolo al menú del principio.

Versión en línea de comandos

Todo inicia con una búsqueda que escribirá el usuario. Debe ingresar una o varias palabras clave:

```
Ingrese su búsqueda: motorola moto g42
```

Luego, le preguntará al usuario si quiere en su búsqueda sólo productos nuevos, usados o ambos a la vez:

```
Seleccione una condición de producto:
```

```
0: Nuevos y Usados
```

```
1: Sólo Nuevos
```

```
2: Sólo Usados
```

```
Ingrese una opción de condición de artículos: 0, 1 o 2: 1
```

En caso de que la búsqueda no disponga de nuevos o de usados, se lo hará saber y a cambio no discriminará por condición:

```
No hay artículos con esa condición en su búsqueda, volviendo a la opción "0" por default...
```

A continuación, le informa de la cantidad de artículos que encontró. En base a este dato, le pregunta al usuario cuántos artículos quiere en su set de datos:

```
Hay exactamente 17 artículos con esa condición.
```

```
Puede traerlos todos o parte de ellos.
```

```
Ingrese 0 si desea todos los artículos ó ingrese una cantidad entre 1 y 17: 0
```

Hace un estimado del tiempo el cual tardará el algoritmo en recabar la cantidad determinada previamente:

```
-----  
Tiempo estimado calculado: 00:00:32  
-----
```

Finalmente, el programa le informa que la planilla con la recolección se depositó en la carpeta 'resultados'. Además, menciona cuántas páginas se visitaron, cuántos artículos se recabaron y el tiempo total de ejecución.

```
Planilla de cálculos "motorola moto g42 - Sólo Nuevos.xlsx" creada correctamente en la carpeta "resultados".  
Páginas visitadas: 1  
Artículos recabados: 17  
Tiempo transcurrido: 00:00:31.16
```

Restricciones

Las restricciones que hay es que sólo funciona para el sitio de ventas de MercadoLibre. Si se quiere adaptar este algoritmo a otro sitio, debe investigarse cómo se disponen los datos y consecutivamente cambiar todas las etiquetas a buscar. Además, el otro sitio debe admitir algoritmos de scraping. Por ejemplo, Amazon no admite scrapers de ningún tipo o, si lo hace, es cada cierto intervalo de tiempo mucho más largo de lo manejado en este proyecto (1 segundo por request).

El tiempo total de ejecución puede variar según su conexión a internet, el tiempo de demora de respuesta del servidor, su hardware, etc.

Realizar búsquedas muy generales provocarán que los registros adquieran columnas que probablemente no les corresponde. Se recomienda fuertemente que las búsquedas sean precisas.

Tecnologías usadas

Python

Python es un lenguaje de programación ampliamente utilizado en las aplicaciones web, el desarrollo de software, la ciencia de datos y el machine learning (ML). Los desarrolladores utilizan Python porque es eficiente y fácil de aprender, además de que se puede ejecutar en muchas plataformas diferentes. Python se puede descargar gratis, se integra bien a todos los tipos de sistemas y aumenta la velocidad del desarrollo.

PyCharm

PyCharm es el IDE para Python desarrollado por JetBrains. Su integración con herramientas colaborativas como GitHub lo convierten en el mejor aliado para el desarrollador y su equipo en este lenguaje. Funciona 'right out of the box' sin necesidad de instalar plugins ni extensiones adicionales.

GitHub

GitHub es una plataforma de desarrollo colaborativo basada en la web que utiliza el sistema de control de versiones Git. El enlace al repositorio del proyecto se encuentra aquí: <https://github.com/cris-andrade-97/MineriaLibre>

Librerías de Python más destacadas del proyecto

Pandas

Pandas es una biblioteca de Python utilizada para el análisis y manipulación de datos. Proporciona estructuras de datos flexibles y eficientes que permiten la manipulación, limpieza y transformación de datos de manera sencilla. Pandas es esencial en el ámbito de la ciencia de datos y análisis estadístico, ya que facilita la importación y exportación de datos desde diversas fuentes.

BeautifulSoup

BeautifulSoup es una popular biblioteca de Python para realizar tareas de web scraping y análisis de datos web, proporcionando una herramienta robusta para extraer datos valiosos de sitios y aplicaciones web. Se puede buscar, filtrar y manipular datos HTML de manera eficiente.

PyQt5

PyQt5 es una librería que nos permite usar interfaces de usuario creadas en Qt Designer al estilo 'click and drag' como lo es también JForm para Java. Luego, con el conversor pyuic5, convertimos nuestra interfaz desde la extensión 'ui' a 'py' para poder manipular en Python.

Lista de todas las librerías usadas en el proyecto

- pandas
- numpy
- bs4
- requests
- time
- warnings
- os
- openpyxl
- PyQt5

Consideraciones finales

Una vez terminado el proyecto, podré decir que he aprendido a manejar lenguajes, librerías y aplicaciones que no conocía anteriormente y que ahora son algo más en mi conocimiento como programador. También aprendí a dar una identidad a lo que uno hace, lo que me deja satisfecho saber que pude lograr crear un explotador de datos dinámico.

Conclusiones

Logros del proyecto

El programa logró satisfactoriamente extraer cada artículo amoldando el set de datos mientras se avanza.

Se manejan los errores presentes en solicitudes a servidor como timeouts y estatus 404.

Se logra predecir correctamente el tiempo estimado de ejecución la mayoría de las veces.

Lecciones aprendidas

Escuchar segundas opiniones desde el inicio del proyecto.

Mejoras para el futuro

Abarcar más filtros de búsqueda, haciendo el dataset resultante aún más personalizable.

Instructivo de instalación

Se confeccionó una serie de instrucciones que cambian según su sistema operativo. Para estos instructivos se seleccionaron dos sistemas operativos. Un sistema libre basado en Linux y un sistema operativo propietario.

El SO libre basado en Linux seleccionado:

- Ubuntu y sus derivados

El SO propietario elegido:

- Microsoft Windows

Sistemas operativos libres basados en Linux

Abra la terminal bash de su sistema operativo y ejecute los comandos en orden:

Ubuntu y sus derivados:

```
sudo apt update
sudo apt install python3 python3-pip git
pip install pandas numpy bs4 pyqt5 openpyxl requests
```

Sistema operativo propietario

Microsoft Windows

Asegúrese de saber qué arquitectura, x64 u x86, es su Windows antes de descargar cualquier software.

- Descargue la última versión de [Git](#)
- Descargue la última versión de [Python3](#)
- Abra la terminal de Windows con privilegios de administrador y escriba los siguientes comandos en orden:

```
curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py
python get-pip.py
pip install pandas numpy bs4 openpyxl pyqt5 requests
```

Ejecutar el programa

El usuario dispone de dos versiones: Una con interfaz gráfica, y otra sólo con línea de comandos.

Versión con interfaz gráfica

En una terminal, indistintamente de su sistema operativo, siga estos pasos:

- Clone el repositorio del proyecto en cualquier carpeta que desee:
`git clone https://github.com/cris-andrade-97/MineriaLibre`
- Ubíquese dentro de la carpeta del proyecto y dentro de la versión con interfaz gráfica:
`cd MineríaLibre/interfaz_grafica`
- Ejecute el programa con:
`python run.py`

Versión en línea de comandos

En una terminal, indistintamente de su sistema operativo, siga estos pasos:

- Clone el repositorio del proyecto en cualquier carpeta que desee:
`git clone https://github.com/cris-andrade-97/MineriaLibre`
- Ubíquese dentro de la carpeta del proyecto y dentro de la versión con línea de comandos:
`cd MineríaLibre/linea_comandos`
- Ejecute el programa con:
`python mineria.py`