MineríaLibre

Integrante:

- 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 los datos.

Abstract

This project called 'MineríaLibre' facilitates data mining on items sold on the Mercado Libre sales site based on any search conducted, discriminating by product condition, whether it's new, used, or both. The dataset adapts in its entirety as the 'scraping' progresses to incorporate each of the collected items. The result is a spreadsheet full of the data.

Introducción

En la era digital actual, la creciente necesidad de datos para análisis y toma de decisiones

ha alcanzado proporciones exponenciales. En este panorama, el dato se ha consolidado como el recurso más valioso, incluso siendo comparado con el petróleo. Este fenómeno ha impulsado la creación de herramientas innovadoras y proyectos ambiciosos que buscan satisfacer esta creciente demanda.

Entonces surge este proyecto llamado 'MineríaLibre', una herramienta diseñada para extraer conocimiento de las vastas cantidades de información alojadas en la plataforma de comercio electrónico MercadoLibre. En un mundo donde cada clic deja un rastro digital, 'MineríaLibre' surge como 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 jocosamente 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 será 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.

Descripción general

En esta sección hablaremos de la descripción general 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.

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:

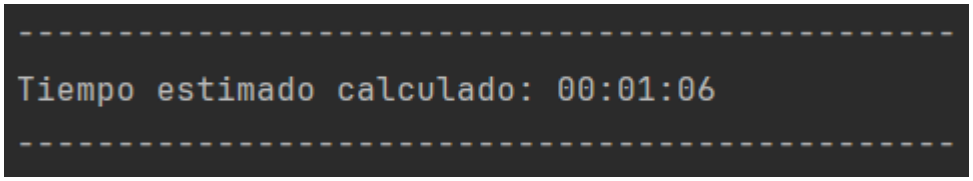
```
No hay artículos con esa condición en su búsqueda, volviendo a la opción "0" por default...
```

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:

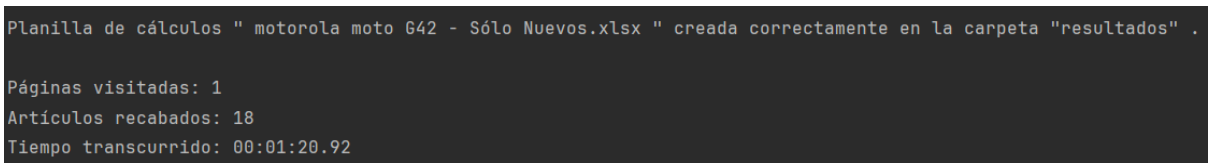
```
Hay exactamente 18 artículos con esa condición.
Puede traerlos todos o parte de ellos.

Ingrese 0 si desea todos los artículos ó ingrese una cantidad entre 1 y 18: 0
```

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



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



Las restricciones que hay para este algoritmo es que sólo funciona para el sitio de ventas de MercadoLibre. Además, el algoritmo debe mantenerse actualizado ya que la forma de ubicar los datos en HTML puede cambiar con el tiempo.

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

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 poderosa 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, así como la realización de operaciones complejas como agrupaciones, fusiones y filtrado.

BeautifulSoup

BeautifulSoup es una popular biblioteca de Python para extraer datos de archivos HTML y XML. Permite analizar documentos web y navegar por su estructura de manera sencilla, facilitando la extracción de información específica. Se puede buscar, filtrar y manipular datos HTML de manera eficiente con sintaxis sencilla. BeautifulSoup es ampliamente utilizada en tareas de web scraping y análisis de datos web, proporcionando una herramienta robusta para extraer datos valiosos de sitios web y aplicaciones web.

Lista de todas las librerías usadas en el proyecto

- pandas
- numpy
- bs4 (BeautifulSoup)
- requests
- time
- warnings
- os
- openpyxl
- odfpy

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.

Instructivo de instalación

Se confeccionó una serie de instrucciones que cambian según su sistema operativo. Para estos instructivos se seleccionaron cuatro sistemas operativos. Tres sistemas libres basados en Linux y un sistema operativo propietario.

Los SO libres basados en Linux seleccionados:

- Ubuntu y sus derivados
- Arch Linux y sus derivados
- Debian

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 odfpy openpyxl
```

Arch Linux y sus derivados:

```
sudo pacman -Syu
sudo pacman -S python python-pip git
pip install pandas numpy bs4 odfpy openpyxl
```

Debian:

```
sudo apt-get update
sudo apt install python3 python3-pip git
pip install pandas numpy bs4 odfpy openpyxl --break-system-packages
```

Sistema operativo propietario

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

Microsoft Windows

- Descargue la última versión de Git
- Descargue la última versión de Python3
- Abra la terminal 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 odfpy openpyxl
```

Ejecutar el programa

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 en la carpeta del proyecto:

```
cd MineriaLibre
```

- Ejecute el programa con alguno de los siguientes comandos:

```
python3 mineria.py
python mineria.py
```