Web scraping en Python (noviembre 2023)

Bohórquez Tejada Nicolás.

Programación para la ciencia de datos.

Universidad Politécnico Grancolombiano.

***Resumen*— En este proyecto de la asignatura "Programación para la Ciencia de Datos", se empleará la técnica de web scraping con el propósito de generar y mantener actualizada información en constante cambio de manera automatizada. Para esta iniciativa, he seleccionado un sitio web de venta de libros nacionales e internacionales, “Busca Libre” como fuente de datos, lo que implica trabajar con diversos tipos de información, además para el proyecto se pide que se trabajen además de los datos cualitativos y cuantitativos, con datos de tipo avanzado, como lo podrían ser imágenes, textos, archivos de audio etc.… El enfoque principal es automatizar la extracción y actualización de información relevante de un sitio web de venta de libros, permitiendo así el análisis descriptivo de datos en tiempo real**

I. Introducción

En este proyecto, exploraremos las capacidades del web scraping para la generación y actualización automatizada de datos, he seleccionado un portal de venta de libros como fuente de datos, un entorno dinámico y en constante evolución que ofrece una amplia gama de información de interés; La elección de esta página web, es para evidenciar y además nos permitirá trabajar con diversos tipos de datos, desde aquellos de naturaleza cualitativa, cuantitativa e información avanzada, como textos completos e imágenes. El objetivo central es demostrar cómo el web scraping puede ser una herramienta valiosa para acceder a información relevante en tiempo real y, a su vez, aplicar técnicas de análisis de datos que se asocian directamente a la carrera de ciencia de datos, para ayuda al web scraping de este proyecto y el análisis, de este, partiremos del libro “Practical Web Scraping for Data Science: Best Practices and Examples with Python” (Autor(es): Seppe vanden Broucke, Bart Baesens, Año: 2018, Editorial: Apress, ISBN-13: 9781484235812)[1] Este libro nos ayudará sobre todo a la planeación del proyecto para darle un enfoque a la ciencia de datos que es a lo que queremos llegar a futuro además que es trabajado en Python y nos ayuda a mirar que opciones y que utilizan para el proceso de web scraping en sus proyectos; ya que otro objetivo es poder aplicar lo trabajado este semestre en clase para extraer conocimientos significativos. Además de hablar de la importancia comercial que tiene “Busca libre” a nivel nacional e internacional, ya que se considera uno de los principales sitios web, venta de libros de todo el mundo, pero además me atrevería a decir que es el portal de venta de libros más grande de latino américa, es por eso que su valor comercial es muy importante, ya que al ser tan reconocido y estar desde hace décadas en lo más arriba de las ventas de diversos tipos de libros, es un gran espacio para poder incluso publicar libros propios para llegar a un mayor alcance . Es importante resaltar que en este documento no se hablará de código literalmente ni se emplearán líneas de programación, lo invito a revisar el repositorio de este proyecto donde podrá encontrar todo el código y el archivo sql de la tabla de la base de datos, hágalo por el siguiente enlace: https://github.com/NicoBTpro/Web-scraping-busca-libre [2] además tengo una explicación breve de 6 min que puede observar por el siguiente enlace: https://youtu.be/aeIF49Bv60Q [3] además, en este mismo canal encontrará una sustentación de mayor duración pero explicada más a fondo.

II. Métodos y materiales

Para entender más a fondo el desarrollo de este proyecto y tener objetivos específicos que cumplir, debemos responder unas preguntas clave que guiará el rumbo de dicho proyecto.

A. URL

La URL básica la cual nos llevará a la pagina a la cual se le realizará el web scraping es la siguiente: https://www.buscalibre.com.co/libros/search

B. Población objetivo.

Como es un sitio de artículos de venta de libros no solamente nacionales (Colombia) sino también internacionales, la población objetivo en este caso será la población colombiana que utiliza esta página para adquirir libros, por lo que la divisa

utilizada será COP, y además, no en todos los países los precios de los libros ni los descuentos son los mismos..

C. *Variables objetivo*

Las variables objetivo son fundamentales para el desarrollo de este proyecto, ya que marcarán el desarrollo de este, por ende, es que se deben de elegir cuidadosamente, después de una investigación [2] hemos encontrado algunas variables para tener en cuenta, además de también incluir las que creo necesarias. Las variables son las siguientes

- Título del libro
- Autor del libro
- Editorial del libro
- Precio en COP del libro
- Precio en COP del descuento del libro
- Porcentaje del descuento del libro
- Stock del libro
- Año de publicación
- URL de la imagen de la portada del libro

D. *Explicación de las variables*

- Título del libro:
  Contendrá el título o encabezado de cada libro. Cualitativa
- Autor del libro:
  Indicará quien publico dicho libro. Cualitativa
- Año de publicación:
  Año exacto en el que se publicó dicho libro. Cuantitativa-Discreta
- Editorial del libro:
  Incluirá el nombre de la editorial que publico el libro. Cualitativa
- URL de la portada del libro:
  Contendrá el enlace directo para observar la caratula de la portada del libro. Cualitativa
- Precio en COP del libro:
  Contendrá el precio en la divisa colombiana del libro. Cuantitativa-continua.
- Precio en COP del descuento del libro:
  Contendrá el precio por el cual se puede adquirir con el descuento aplicado. Cuantitativa-continua
- Porcentaje del descuento:
  Contendrá el valor numérico del descuento. Cuantitativa-discreta
- Stock del libro: contendrá el numero de libros disponibles que quedan. Cuantitativa-discreta

III. Planteamiento del código en python

Para este apartado, me guiaré en principio del primer código de web scraping que hicimos en clase, después de esto, comenzaré por crear mi script en postgresql, donde se almacenarán mis datos recolectados, luego, se realizará el scraping para los n libros de los n géneros que yo quiera estudiar, luego comenzaré a realizar diferentes filtro y consultas para poder trabajar con los datos desde mi cuaderno de python, luego mediante un ciclo y funciones poder descargar los archivos de imagen en una carpeta local la cual será seleccionada por una interfaz de usuario, después de esto, se analizarán las imágenes recolectada de las portadas de los libros para identificar los colores principales de estas, también habrá una función que permita desde un parámetro que se le pasa a una función, filtrar por el porcentaje de descuento de los libros, por ejemplo si quiero ver o consultar solo los libros que tengan un descuento igual o mayor al 30, este será el valor de la variable que se le pasará a la función, luego pasaremos a la parte final del código que será realizar los distintos tipos de graficas entre las variables que considere necesarias.

IV. Planteamiento de la base de datos

Se creará una tabla para almacenar todos los datos recolectados por el programa, esta tabla tendrá las siguientes variables y tipos de dato (literalmente esta es la tabla):

```
CREATE TABLE Libros (
    id SERIAL PRIMARY KEY,
    titulo VARCHAR(255) NOT NULL,
    autor VARCHAR(255) NOT NULL,
    editorial VARCHAR(255) NOT NULL,
    anio_publicacion INTEGER NOT NULL,
    stock INTEGER NOT NULL,
    descuento_porcentaje INTEGER NOT NULL,
    precio_antes DECIMAL(10, 2) NOT NULL,
    precio_ahora DECIMAL(10, 2) NOT NULL,
    imagen_url VARCHAR(255)
);
```

Sabiendo los nombres de las variables y el tipo de dato que requieren, ya es simplemente hacer el casting o la conversión implícita al tipo de dato determinado en nuestro programa y luego hacer la inserción en esta base de datos mediante una conexión desde python

V. Planteamiento del problema

Este proyecto se plantea con la intención de responder a varias preguntas de investigación importantes:

- Optimización de Precios:
  ¿Cómo podemos identificar y analizar libros con descuentos significativos para ofrecer a los usuarios oportunidades de compra atractivas?
- Popularidad de Autores y Editoriales:
  ¿Existen autores o editoriales cuyos libros son más populares entre los compradores?
- Impacto de la Imagen de Portada:
  ¿Podremos saber cuáles son los colores que más predominan en las portadas de los libros más vendidos?

- Graficas, ¿qué podemos interpretar de las diferentes graficas realizadas?

VI. Resultados y discusión.

Antes de comenzar a hablar de los resultados de discusión, para no hacer extenso el documento ni tedioso de leer por tantas partes de código además de extensas, recortadas por el formato, lo invito a que usted mismo mire el proyecto en el siguiente enlace: https://github.com/NicoBTpro/Web-scraping-busca-libre [2] de igual forma el video está disponible mediante el siguiente vinculo: https://youtu.be/aeIF49Bv60Q [3]

- Optimización de precios:
  Para este apartado, diseñé una función para que el programa después de hacer el web scraping a los n libros de las n secciones de nuestro interés, recolectara información valiosa que usted puede encontrar en el apartado de construcción de la base de datos anteriormente en este documento, donde después de establecer la conexión con la base de datos, la filtramos por un parámetro descrito, en el que el parámetro es el porcentaje de descuento que queremos establecer, y así, arrojándonos todos los resultados de acorde a lo que queremos, además, en la parte final del código donde se encuentran las gráficas, podemos ver un grafico de torta en el cual podemos observar los autores (10) con mayor descuento en sus libros.
- Popularidad de Autores y Editoriales:
  Lastimosamente para este apartado la respuesta no es clara, ya que busca libre no nos da la información acerca de cuantos libros ha vendido ese actor ni cuantas copias de un libro en especifico hay vendidos, pero si nos muestran el stock de cada libro, entonces para este apartado se comparó el stock de los libros de cada autor, siendo así el autor con más stock se infiere que es el autor con mayor demanda para sus libros, por ende, el que mayor popularidad tiene, de igual forma, este apartado cuenta con un grafico de barras donde se pueden observar los autores con mayor stock y por ende con mayor demanda.
- Impacto de la Imagen de Portada
  Este apartado se trabajo con el tipo de dato avanzado, la imagen; lo primero, como se mencionó en “la construcción de la base de datos” tenemos una variable que almacena las URL de las imágenes de portada de dichos libros, lo que se hizo fue establecer la conexión con la base de datos, hacer un ciclo que recorriera esa variable y fuera almacenando las imágenes en una carpeta que mediante una interfaz se le muestra al usuario el cual decide en qué carpeta guardarlas, de este modo, habiendo guardado las n imágenes de los n libros, en una carpeta, lo cual nos ayuda para el siguiente paso
- Determinar los colores principales de las portadas:
  Una vez descargadas las imágenes y guardadas en la carpeta, con el path de dicha carpeta se la pasaremos a una función, la cual recorre los elementos de esta carpeta y analiza imagen por imagen determinando su color principal, para al final, determinar los 3 colores principales que mas se repiten en todos los libros y guardar estos 3 colores en una carpeta o directorio que le coloquemos a esta función.
- Graficas
  Las distintas graficas que se pueden observar en el documento son hechas con matplotlib se saca los datos a graficar directamente de la base de datos y se trabajan sobre estos mismos.

VII. Conclusiones.

Este proyecto de web scraping centrado en el portal de venta de libros "Busca Libre" ha sido una exploración integral de las capacidades del web scraping y el análisis de datos. La elección de esta plataforma proporcionó una fuente diversa de información sobre libros, desde datos cuantitativos como precios y descuentos hasta información cualitativa como autores, editoriales y títulos.

La estructura del proyecto abordó metódicamente aspectos clave, desde la definición de la URL objetivo y la población objetivo hasta la identificación de variables importantes para recopilar. El planteamiento del código en Python siguió un enfoque sistemático, desde el scraping inicial hasta la creación de gráficos y análisis detallados. Además, se estableció una base de datos en PostgreSQL para almacenar y gestionar eficientemente los datos recolectados.

Los resultados y discusiones se enfocaron en preguntas clave de investigación, como la optimización de precios, la popularidad de autores y editoriales, y el impacto de las imágenes de portada. A pesar de ciertas limitaciones en la disponibilidad de datos específicos, se lograron aspectos valiosos, como la identificación de libros con descuentos significativos y la determinación de autores populares basados en el stock.

La capacidad para extraer información de imágenes, como la identificación de colores predominantes en las portadas de libros, agregó un componente visual y estético al análisis de datos. Las visualizaciones, realizadas mediante gráficos utilizando matplotlib, proporcionaron una representación clara y accesible de los resultados.

En conclusión, este proyecto no solo demostró la viabilidad y la utilidad del web scraping para obtener datos en tiempo real, sino que también destacó cómo estas técnicas pueden integrarse con el análisis de datos y la creación de visualizaciones significativas los invito a revisar no solo el repositorio creado para este proyecto en Git hub: https://github.com/NicoBTpro/Web-scraping-busca-libre [2] , sino a también ver el video publicado donde explico más a fondo este programa: https://youtu.be/aeIF49Bv60Q [3] y en el mismo canal de youtube encontrará un video de la

explicación con mayor duración pero más detallado; sin más que decir, este es el final del proyecto, gracias por su atención y espero al menos les haya parecido interesante este proyecto llevado a cabo. Nicolás Bohórquez Tejada

REFERENCIAS

[1] AUTOR(ES): SEPPE VANDEN BROUCKE, BART BAESENS, “PRACTICAL WEB SCRAPING FOR DATA SCIENCE: BEST PRACTICES AND EXAMPLES WITH PYTHON”, EDITORIAL APRESS. ISBN-13: 9781484235812

[2] Nicolás Bohórquez Tejada “Web-scraping-busca-libre” 30 de noviembre del año 2023, disponible en https://github.com/NicoBTpro/Web-scraping-busca-libre

[3] Nicolás Bohórquez Tejada” Web scraping busca libre” 30 de noviembre del año 2023, video en línea disponible en: https://youtu.be/aeIF49Bv60Q