Manual de Usuario del Crawler Web

1. Introducción

El presente manual tiene como objetivo proporcionar a los usuarios una guía detallada sobre el uso del crawler web desarrollado, el cual tiene como finalidad agilizar el proceso de pentest ético al identificar enlaces, tecnologías y otros datos relevantes en un sitio web objetivo. Mediante la exploración automatizada y el análisis exhaustivo, se busca mejorar la eficiencia y precisión del proceso de evaluación de seguridad.

1.1. Objetivos del Crawler Web

El objetivo principal del proyecto es desarrollar un crawler web eficiente que cumpla con los siguientes objetivos generales:

- Desarrollar un crawler web capaz de explorar un sitio web objetivo y extraer información relevante.
- Implementar la detección de tecnologías y frameworks empleados en el sitio web para identificar posibles vulnerabilidades asociadas.
- Exportar los resultados en diferentes formatos para facilitar el análisis posterior y la integración con otras herramientas.
- Generar automáticamente un informe PDF que presente los resultados de manera clara y organizada.
- Realizar pruebas en entornos controlados para validar el funcionamiento del crawler y garantizar su seguridad y eficacia.

1.2. Audiencia

Este manual está dirigido a profesionales y entusiastas del pentest ético que deseen utilizar el crawler web como una herramienta eficiente para identificar y evaluar vulnerabilidades en sistemas y aplicaciones web.

2. Requisitos del sistema

Asegúrese de cumplir con los siguientes requisitos del sistema antes de utilizar el crawler web:

- Sistema operativo compatible: Windows.
- Versión de Python requerida: 3.11.4 o compatibles.
- Otras dependencias o bibliotecas requeridas:
    - scrapy
    - csv
    - builtwith
    - re
    - json
    - os
    - datetime
    - scrapy.exporters
    - reportlab.lib.pagesizes
    - reportlab.pdfgen

3. Instalación y configuración

Sigue los pasos a continuación para instalar y configurar el crawler web:

- Instalación de Python: Descarga e instala Python 3.11.4 desde el sitio web oficial de Python (https://www.python.org).
- Instalación de dependencias: Abre una ventana de comandos y ejecuta el siguiente comando para instalar las dependencias necesarias:

```
Comando: Instalación de Scrapy desde CMD, este comando descargará e instalará Scrapy con las dependencias necesarias.
pip install scrapy
```

Además, asegúrate de tener instaladas las bibliotecas adicionales mencionadas anteriormente.

Una vez instalado Scrapy se podrá crear un nuevo proyecto con el siguiente comando:

```
Scrapy stratproject “nombre_proyeto”
```

- Descarga del crawler web: Descarga el archivo del crawler web desde el repositorio https://github.com/Vanessa1114/Crawler-Web y guárdelo en la ruta del proyecto creado.
- Configuración del crawler: Abre el archivo de configuración del crawler y establece la URL de inicio del sitio web objetivo. Puedes modificar otras configuraciones según tus necesidades.
- Para facilitar la edición del crawler web, se recomienda utilizar un entorno de desarrollo integrado (IDE) de Python, como PyCharm o Visual Studio Code. Estos IDEs proporcionan herramientas y características adicionales para una programación más eficiente y una mejor experiencia de desarrollo.

4. Uso básico del crawler web

El crawler web se utiliza desde la línea de comandos. A continuación, se muestra una descripción del comando utilizado para la ejecución del crawler:

```
Scrapy runspider .\nombre_del_crawler.py
```

Descripción: Inicia el crawler web y comienza a explorar el sitio web objetivo.

Nota: Asegúrate de ubicarte en el directorio donde se encuentra el archivo del crawler web antes de ejecutar el comando.

5. Resultados y exportación de datos

El crawler web generará resultados y exportará los datos extraídos en varios formatos. A continuación, se describen los pasos para acceder a los resultados y exportar los datos:

Carpeta de resultados: Después de ejecutar el crawler web, se creará una carpeta con la fecha actual en el directorio de trabajo. Dentro de esta carpeta se guardarán los archivos de resultados.

Archivos de resultados:

- output.csv: Archivo CSV que contiene los enlaces, títulos y tecnologías encontradas en el sitio web objetivo.
- output.json: Archivo JSON que contiene los resultados en formato JSON para un análisis posterior.
- output.xml: Archivo XML que exporta los resultados en un formato estructurado.
- output.pdf: Informe PDF generado automáticamente que presenta los resultados de manera clara y organizada.

6. Personalización y configuración avanzada

El crawler web puede ser personalizado y configurado según sus necesidades específicas. Algunas opciones de personalización y configuración avanzadas:

- Configuración de campos de salida: Puedes modificar los campos de salida en el archivo del crawler web para incluir información adicional relevante.
- Filtrado de enlaces: Puedes agregar filtros y reglas específicas para evitar la exploración de ciertos enlaces o páginas.

7. Actualizaciones y mejoras futuras

El crawler web está en constante desarrollo y mejora. Mantente actualizado sobre las últimas actualizaciones y mejoras visitando el repositorio https://github.com/Vanessa1114/Crawler-Web.