Documentación de la Actividad

Introducción

El objetivo de esta actividad, correspondiente a la **ACTIVIDAD 1.2**, consistió en desarrollar un script de **web scraping** utilizando las herramientas requests y BeautifulSoup. El propósito principal fue realizar la recolección de información desde Wikipedia mediante la búsqueda de un término, obteniendo el título, la descripción y la primera imagen de la página correspondiente.

El ejercicio también incorpora pruebas unitarias como un agregado adicional, con el fin de validar el correcto funcionamiento del script para diversas búsquedas.

Desarrollo de la Actividad

Herramientas y Tecnologías

Para realizar la actividad se utilizaron las siguientes tecnologías:

- **Lenguaje de programación:** Python
- **Librerías principales:**
  - requests: Para realizar solicitudes HTTP a Wikipedia.
  - BeautifulSoup: Para analizar y extraer datos del HTML devuelto por Wikipedia.
  - PIL (Python Imaging Library): Para manejar y mostrar imágenes descargadas.
  - **Herramienta opcional agregada:** pytest para ejecutar pruebas unitarias.

Estructura del Script

1. **Solicitud a Wikipedia:**
   - Se desarrolló una función para realizar una solicitud HTTP a la página de Wikipedia correspondiente al término de búsqueda ingresado. En caso de que la búsqueda sea ambigua o lleve a una página de desambiguación, el script selecciona automáticamente el primer resultado relevante.
2. **Análisis de la Página:**
   - Usando BeautifulSoup, se analizó el HTML de la página devuelta. Se extrajeron los elementos más importantes: el título de la página, el primer párrafo de texto que contenga una descripción significativa, y la primera imagen dentro de la infobox (si está disponible).
3. **Visualización de la Imagen:**
   - Si se encuentra una imagen en la página, el script descarga y muestra la imagen usando PIL.
4. **Ejecución del Script:**

- El script puede ejecutarse desde la línea de comandos con un término de búsqueda como argumento, o bien solicitar el término al usuario si no se pasa ningún argumento.

Pruebas Unitarias (Agregado)

Además del script de scraping, se desarrollaron pruebas unitarias como una mejora adicional. Estas pruebas usan pytest para verificar el correcto funcionamiento del script con palabras clave populares, como:

- **"pokemon"**
- **"mario"**
- **"warcraft"**
- **"legend of zelda"**

Las pruebas validan que se obtiene un título, una descripción relevante y una URL de imagen válida para cada una de estas palabras clave.

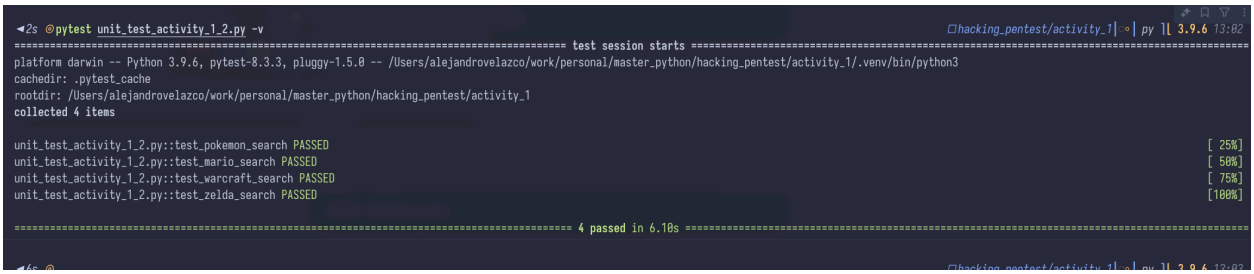
Capturas de Pantalla

1. Ejecución del Script con "Super Mario"

- Captura que muestra la ejecución del script al buscar el término "Super Mario", incluyendo el título, descripción, y la URL de la imagen.

2. Resultados de las Pruebas Unitarias

- Captura de los resultados de la ejecución de las pruebas unitarias con pytest, mostrando la validación de las búsquedas de las palabras "pokemon", "mario", "warcraft", y "legend of zelda".



Conclusiones

El desarrollo de esta actividad permitió poner en práctica las técnicas de **web scraping** con requests y BeautifulSoup, logrando automatizar la recolección de información desde Wikipedia. Adicionalmente, la visualización de imágenes mediante PIL proporcionó una interfaz simple para mostrar resultados más completos al usuario.

Las pruebas unitarias implementadas, aunque no eran requeridas en el objetivo original de la actividad, permiten validar que el script se comporta adecuadamente en distintos escenarios, contribuyendo a una mayor robustez y fiabilidad del código.

En conclusión, este ejercicio ha sido útil para entender el funcionamiento del scraping de datos web y la importancia de implementar pruebas unitarias para asegurar la correcta ejecución de un programa en distintos entornos.