

Documento Explicativo del Código

Este código está diseñado para descargar archivos PDF de un sitio web automáticamente. A continuación, te explico de forma sencilla qué hace cada parte del código.

¿Qué Hace el Código?

1. **Crear una carpeta para los archivos descargados:** El código comienza creando una carpeta llamada "downloads" en la misma ubicación donde se encuentra el programa. Si esta carpeta no existe, se crea automáticamente para almacenar los archivos PDF descargados.
2. **Abrir un navegador web de manera automática:** Utiliza una herramienta llamada **Selenium** para abrir un navegador web, en este caso **Google Chrome**. En lugar de abrir el navegador manualmente, el código lo hace de manera automática. Esto permite que el programa navegue por internet sin intervención humana.
3. **Buscar los enlaces de las secciones que contienen archivos PDF:** Una vez que se abre la página, el código busca los enlaces dentro de la página web que tienen una parte específica en la URL. Estos enlaces apuntan a archivos PDF que el programa debe descargar. El código se asegura de no seleccionar enlaces que ya contengan la extensión ".pdf" en su URL, asegurando que solo se descarguen los archivos correctos.
4. **Generar y descargar los archivos PDF:** Después de obtener los enlaces, el código crea la URL final del archivo PDF agregando la extensión ".pdf" al final de cada enlace encontrado. Luego, se descarga el archivo en el directorio "downloads".
5. **Guardar los archivos con un nombre apropiado:** El código guarda cada archivo PDF descargado con un nombre basado en el nombre del archivo de la URL. Esto facilita la organización de los archivos y asegura que cada uno tenga un nombre único y claro.

Librerías Usadas

Selenium (WebDriver y ChromeDriver):

- **Selenium** es una herramienta para automatizar la interacción con navegadores web. Permite abrir, navegar, y realizar búsquedas en sitios web de manera programática.
- **WebDriver** es la interfaz que interactúa con el navegador, en este caso, Chrome a través del **ChromeDriver**.
- Con Selenium, el programa puede simular un comportamiento de usuario real (como hacer clic en enlaces) para obtener información de la página web.

Apache Commons IO (FileUtils):

- **Apache Commons IO** es una librería de utilidades para facilitar la manipulación de archivos y flujos de datos. En este caso, se usa la clase **FileUtils** para descargar los archivos PDF desde una URL y guardarlos en el sistema de archivos local.
- El método `FileUtils.copyURLToFile()` permite descargar un archivo de una URL y guardarlo directamente en una ubicación especificada.

¿Cómo Funciona ?

1. **Configura el directorio de descarga:** El primer paso es configurar el directorio donde se guardarán los archivos descargados. Si la carpeta "downloads" no existe, se crea automáticamente. Esto asegura que todos los archivos se guarden en un lugar específico y organizado dentro del sistema.
2. **Inicia el navegador web:** Usando Selenium, el programa inicia un navegador Google Chrome (mediante ChromeDriver) y accede a la página web especificada. Selenium permite simular la navegación como si un usuario estuviera abriendo manualmente el navegador y buscando los archivos.
3. **Busca los enlaces de los PDFs:** El código utiliza una expresión XPath para identificar los enlaces que contienen la palabra "md/en/" pero no terminan con ".pdf". XPath es una herramienta poderosa para navegar por la estructura HTML de una página web, permitiendo localizar elementos como enlaces, imágenes, formularios, etc. Con esta búsqueda, el código asegura que solo los enlaces a las secciones correctas sean seleccionados.
4. **Genera y descarga los archivos PDF:** Una vez identificados los enlaces, el código construye una URL para cada archivo PDF, agregando ".pdf" al final. Esto genera el enlace completo al archivo PDF, basado en la URL de la sección original. Luego, se utiliza `FileUtils.copyURLToFile()` para descargar cada archivo PDF desde la web y guardarlo en la carpeta de "downloads".

Ejemplo de Código Explicado

1. Abrir una página web con Selenium:

Este código muestra cómo usar Selenium para abrir una página web.

```
java Copiar

import org.openqa.selenium.WebDriver;
import org.openqa.selenium.chrome.ChromeDriver;

public class AbrirPagina {
    public static void main(String[] args) {
        // Configurar el path del ChromeDriver
        System.setProperty("webdriver.chrome.driver", "path/to/chromedriver");

        // Iniciar el navegador
        WebDriver driver = new ChromeDriver();

        // Acceder a la página web
        driver.get("https://nachoiborraies.github.io/java/");

        // Espera 2 segundos antes de cerrar el navegador
        try {
            Thread.sleep(2000);
        } catch (InterruptedException e) {
            e.printStackTrace();
        }

        // Cerrar el navegador
        driver.quit();
    }
}
```

¿Qué hace este código?

- Configura el navegador Chrome para abrirlo automáticamente.
- Va a la página web que se especifica (<https://nachoiborraies.github.io/java/>).
- Espera 2 segundos para que la página se cargue, y luego cierra el navegador.

2. Descargar un archivo PDF usando **FileUtils** de Apache:

Este fragmento muestra cómo descargar un archivo desde una URL usando **FileUtils.copyURLToFile()**.

```
java Copiar

import org.apache.commons.io.FileUtils;
import java.io.File;
import java.io.IOException;
import java.net.URL;

public class DescargarArchivo {
    public static void main(String[] args) {
        // URL del archivo que quieres descargar
        String archivoURL = "https://nachoiborraies.github.io/java/sample.pdf";

        // Ruta donde se guardará el archivo
        String rutaArchivo = "downloads/sample.pdf";

        // Descargar el archivo
        try {
            FileUtils.copyURLToFile(new URL(archivoURL), new File(rutaArchivo));
            System.out.println("Archivo descargado con éxito");
        } catch (IOException e) {
            System.err.println("Error al descargar el archivo: " + e.getMessage());
        }
    }
}
```

¿Qué hace este código?

- Define la URL de un archivo PDF (<https://nachoiborraies.github.io/java/sample.pdf>).
- Define la ruta donde se guardará el archivo descargado ([downloads/sample.pdf](#)).
- Usa `FileUtils.copyURLToFile()` para descargar el archivo y guardarlo en el lugar indicado.

Conclusión

Al haber ejecutado tendríamos que tener descargados todos los pdf de la página.

downloads

Este equipo

Disco local (C:)

entornos

java

downloads

Buscar en downloads

Nuevo

Cortar

Copiar

Compartir

Eliminar

Ordenar

Ver

Detalles

entornos

comandos git

GitExercises

java

idea

downloads

ejercicios

out

POO

Selenium

lenguaje de marcas

mareo

MSI

Nueva carpeta

oraclexe

PerfLogs

programacion

ProgramData

sistemas

Usuarios

visio

01a.pdf

Microsoft Edge PDF Document

139 KB

01b.pdf

Microsoft Edge PDF Document

171 KB

01c.pdf

Microsoft Edge PDF Document

185 KB

01d.pdf

Microsoft Edge PDF Document

169 KB

01e.pdf

Microsoft Edge PDF Document

167 KB

02a.pdf

Microsoft Edge PDF Document

141 KB

02b.pdf

Microsoft Edge PDF Document

2.00 MB

02c.pdf

Microsoft Edge PDF Document

126 KB

02d.pdf

Microsoft Edge PDF Document

573 KB

03a.pdf

Microsoft Edge PDF Document

160 KB

03b.pdf

Microsoft Edge PDF Document

118 KB

03c.pdf

Microsoft Edge PDF Document

176 KB

03d.pdf

Microsoft Edge PDF Document

165 KB

04a.pdf

Microsoft Edge PDF Document

180 KB

04b.pdf

Microsoft Edge PDF Document

160 KB

04c.pdf

Microsoft Edge PDF Document

173 KB

04d.pdf

Microsoft Edge PDF Document

225 KB

05a.pdf

Microsoft Edge PDF Document

320 KB

05b.pdf

Microsoft Edge PDF Document

166 KB

05d.pdf

Microsoft Edge PDF Document

128 KB

05e.pdf

Microsoft Edge PDF Document

230 KB

05f.pdf

Microsoft Edge PDF Document

193 KB

06a.pdf

Microsoft Edge PDF Document

122 KB

06b.pdf

Microsoft Edge PDF Document

597 KB

06c.pdf

Microsoft Edge PDF Document

193 KB

06d.pdf

Microsoft Edge PDF Document

138 KB

07a.pdf

Microsoft Edge PDF Document

184 KB

07b.pdf

Microsoft Edge PDF Document

143 KB

07c.pdf

Microsoft Edge PDF Document

140 KB

07d.pdf

Microsoft Edge PDF Document

129 KB

08a.pdf

Microsoft Edge PDF Document

119 KB

08b.pdf

Microsoft Edge PDF Document

141 KB

08c.pdf

Microsoft Edge PDF Document

137 KB

09a.pdf

Microsoft Edge PDF Document

144 KB

09b.pdf

Microsoft Edge PDF Document

125 KB

09c.pdf

Microsoft Edge PDF Document

417 KB

10a.pdf

Microsoft Edge PDF Document

441 KB

10b.pdf

Microsoft Edge PDF Document

202 KB

10c.pdf

Microsoft Edge PDF Document

369 KB

11a.pdf

Microsoft Edge PDF Document

128 KB

11b.pdf

Microsoft Edge PDF Document

175 KB

11c.pdf

Microsoft Edge PDF Document

223 KB

13a.pdf

Microsoft Edge PDF Document

344 KB

13b.pdf

Microsoft Edge PDF Document

563 KB

13c.pdf

Microsoft Edge PDF Document

1.41 MB

14a.pdf

Microsoft Edge PDF Document

957 KB

14b.pdf

Microsoft Edge PDF Document

228 KB

14c.pdf

Microsoft Edge PDF Document

566 KB

14d.pdf

Microsoft Edge PDF Document

395 KB

14e.pdf

Microsoft Edge PDF Document

274 KB

14f.pdf

Microsoft Edge PDF Document

221 KB

14g.pdf

Microsoft Edge PDF Document

205 KB

15a.pdf

Microsoft Edge PDF Document

385 KB

15b.pdf

Microsoft Edge PDF Document

383 KB

54 elementos