**HTML FILTER**

**[ WITH JAVA]**

**This java code accepts an HTML file (from the given absolute or relative location) as an input and do sanitizing on script related elements and attributes and then gives HTML file (at the given location) which contain the filtered source code.**

**CLASS 1: Main class [ contains main method ]**

**import java.io.File;**

**import java.io.IOException;**

***public* class Main {**

***public* *static* void main(String[] args) throws IOException {**

**File inputHtmlFile = new File("C:\\Users\\hp pavilion\\Documents\\GitHub\\Fundamentals-of-software-security-prooject\\sample html document\\OriginalHtmlCode.html");**

**File outputHtmlFile = new File("C:\\Users\\hp pavilion\\Documents\\GitHub\\Fundamentals-of-software-security-prooject\\sample html document\\FilteredHtmlCode.html");**

**HtmlFilterTwo.sanitizeHtmlFile(inputHtmlFile, outputHtmlFile);**

**}**

**}**

**CLASS 2: contain the sanitizing method**

**import java.io.File;**

**import java.io.FileWriter;**

**import java.io.IOException;**

**import org.jsoup.Jsoup;**

**import org.jsoup.nodes.Document;**

**import org.jsoup.nodes.Element;**

**import org.jsoup.select.Elements;**

***public* class HtmlFilterTwo {**

***public* *static* void sanitizeHtmlFile(File inputHtmlFile, File outputHtmlFile) throws IOException {**

***// Parse the input HTML file using Jsoup***

**Document doc = Jsoup.parse(inputHtmlFile, "UTF-8");**

***// Remove all script tags and inline scripts***

**doc.select("script").remove();**

***// Select all elements in the document***

**Elements allElements = doc.select("\*");**

***// Loop over all elements in the document***

**for (Element element : allElements) {**

***// Loop over all attributes of the current element***

**for (org.jsoup.nodes.Attribute attr : element.attributes()) {**

***// Check if the name of the current attribute starts with "on"***

**if (attr.getKey().toLowerCase().startsWith("on")) {**

***// Remove the current attribute from the element if it is an event handler***

**element.removeAttr(attr.getKey());**

**}**

**}**

**}**

***// Write the sanitized HTML to the output file***

**FileWriter writer = new FileWriter(outputHtmlFile);**

**writer.write(doc.html());**

**writer.close();**

**}**

**}**

**EXPLANATION**

**FOR CLASS *HtmlFilterTwo***

* ***Document doc = Jsoup.parse(inputHtmlFile, "UTF-8");***

This line uses the Jsoup library to parse the input HTML file into a Document object. The parse() method takes the input HTML file and the character encoding as arguments.

* ***doc.select("script").remove();***

This line selects all <script> elements in the document using a CSS selector and removes them from the document.

* ***Elements allElements = doc.select("\*");***

This line selects all elements in the document using the \* selector and assigns them to the allElements variable.

* ***for (Element element : allElements) {***

This line starts a loop over all elements in the document.

* ***for (org.jsoup.nodes.Attribute attr : element.attributes()) {***

This line starts a nested loop over all attributes of the current element.

* ***if (attr.getKey().toLowerCase().startsWith("on")) {***

This line checks if the name of the current attribute starts with "on", which is the convention for event handler attributes in HTML.

* ***element.removeAttr(attr.getKey());***

This line removes the current attribute from the element if it is an event handler.

* ***FileWriter writer = new FileWriter(outputHtmlFile);***

This line creates a FileWriter object to write the sanitized HTML to the output file.

* ***writer.write(doc.html());***

This line writes the HTML content of the Document object to the output file using the FileWriter object.

* ***writer.close();***

This line closes the FileWriter object to free up system resources.