Class Overview

- **WebCrawlerII**
    - Class Summary
        - A class focuses on gaining and exporting content of the website given in a URL.
    - Nested Class Summary
        - ImageUrl
    - Constructor Summary

Modifier	Constructor and Description
public	WebCrawlerII(URL url) Construct by a URL and require the content and the links of the website using default method.
public	WebCrawlerII(URL url, String contentMethod) Construct by a URL and require the content of website using given method.
public	WebCrawlerII(URL url, String contentMethod, String urlMethod) Construct by a URL and require the content and links of website using given method.
public	WebCrawlerII(String urlString) Construct by a string and require the content and the links of the website using default method.
public	WebCrawlerII(URL url, String contentMethod) Construct by a string and require the content of website using given method.
public	WebCrawlerII(URL url, String contentMethod, String urlMethod) Construct by a URL and require the content and links of website using given method.

- Element Summary

Modifier	Element and Description
public static String	*USE_BUILT_IN* `Use built-in library`
public static String	*USE_HTTP_CLIENT* `Use the third party library - HttpClient`
public static String	*USE_STRING_PARSER* `Parse in Regex.`
public static String	*USE_SOUP_STRING* `Parse in Jsoup`
public static String	*FILE_DEFAULT_NAME* `Define the default name of the file to be written.`
public static int	*IMAGE_AVERAGE_SIZE* `Denote the average size.`
public static int	*IMAGE_MAX_SIZE* `Define the maximum size.`
private URL	url `The url of the website.`
private String	content `The content(or source code) of the website.`
private Vector<URL>	linkList `The collection of links in the given website.`
private Vector<ImageUrl>	imageList `The collection of URLs of images in the given website.`
int	sizeAvg `The average size of all the images in the given website.`

- Method Summary

Modifier and Type	Method and Description

public void	getContentBuiltIn() Require the content using HttpURLConnection.
public void	getContentHttpClient() Require the content using HttpClent.
public void	getLinkListByString() Parse the links in regex.
public void	getLinkListBySoup() Parse the liinks in Jsoup.
public int	getImageSize(URL aUrl) Get the size of the image given in URL.
public void	getImagesByString() Parse the URLs of images in regex.
public void	getImagesBySoup() Parse the URLS of inages in Jsoup.
public String	getContent() Return the content.
public Vector<URL>	getLinkList() Return the collection of links.
public Vector<URL>	getImageList(int minSize, int maxSize) Return the images whose size is between minSize and maxSize.
public Vector<URL>	getImageList(int minSize) Return the images whose size is above minSize.
public Vector<URL>	getImageList() Return all the images. Return the images whose size is above minSize.
public void	exportImageInFile() Export images.
public void	exportImageInHtml() Export images in Html format.
public void	printTextInFile(String fileName, String fileContent) Write fileContent into a file named fileName.

- **WebContentCrawlerII**
    - Class Summary

    - A class extends ***WebCrawlerII*** and focuses on gaining and exporting the content in the **main body** of website given in a URL.

- Constructor Summary

Modifier	**Constructor and Description**
public	WebContentCrawlerII(URL url)
public	WebContentCrawlerII(URL url, String contentMethod)
public	WebContentCrawlerII(URL url, String contentMethod, String urlMethod)
public	WebContentCrawlerII(String urlString)
public	WebContentCrawlerII(String urlString, String contentMethod)
public	WebContentCrawlerII(String urlString, String contentMethod, String urlMethod)

- Element Summary

Modifer and Type	**Element and Description**
String	text `The text in the main body.`
String	textInHtml `The text organized in Html format in the main body`

- Method Summary

Modifer and Type	**Element and Description**

public void	getContentBuiltIn()
	`Require the content in the main body of website using HttpURLConnection.`

public void	getContentHttpClient()
	`Require the content in the main body of website using HttpClient.`

public void	filterText()
	`Filter the content of main body from original content.`

public String	getContentText()
	`Return the text in the main body.`

public void	printTextInHtml(String fileName) Export the text of main body in Html into a file named fileName.
public void	printTextInPlainText(String fileName) Export the text of main body into a file named fileName.
public void	printTextInHtml() Export the text of main body in Html into a file using a default name.
public void	printTextInPlainText() Export the text of main body into a file using a default name.

Example

- Code

```
aUrl = new URL("http://www.ifanr.com/458506");
WebContentCrawlerII crawler = new(WebContentCrawlerII(aUrl,
WebCrawlerII.USE_HTTP_CLIENT, WebCrawlerII.USE_SOUP_PARSER);
for (URL link : crawler.getLinkList()) {
        System.out.println(link.toString());
}           System.out.println(crawler.getContentText());
crawler.exportImageInHtml();
crawler.exportImageInFile();
crawler.printTextInHtml();
crawler.printTextInPlainText();
```

- Expected Result

```
.
+-- txt
|   +-- WebContent.txt
|   +-- WebContent.html
+-- img
|   +-- 37.jpg
|   +-- 210.jpg
|   +-- D1.jpg
```

```
    |   +-- images.html
```

> Written with StackEdit.