Comp533: Project 2 model description

Frank Wu 260580792

Chin Wang Cheong 260807317

Note

- Due to issues of Touchcore importing Java library class as implementation class, We decide to simulate implementation class by creating class with name in the form of "ImpXXX" to indicate that it is a imported Java library class.

Page Getter

In this design model, we use Java library HtmlPage, UrlValidator and WebClient to obtain html source code by input an url and we also validate the url format before getting html code.

In Page Getter class, getWebPage() will first validate url using isFormatValid() to validate the url format using Java library class UrlValidator which will do check on "http","https" and "ftp". Then an instance of java library class HtmlPage which contains the corresponding Html source code of the page will be returned.

Page Content Processing

In this design model, we use JavaLibrary HtmlElement and String (we create this class because we want to obtain string length which is not support by TouchCore String type). The User's query contains the tagId and xmlPath of the HtmlElement. We will use getHtmlElement() in ContentProcessor class to obtain user's desired content by matching tagId or xmlPath or both.

ContentProcessor and ContentType are partial classes. ContentProcessor's getContent() method is a partial method which will be defined by user. User can customize the format of the output content. The output content will be ContentType class which will be defined by user.

TextContentProcessor will return text content based on user input regular expression. MediaContentProcessor will download the media to the user-specified file location.

Authentication Handling

In this design model, before the ContentProcessor obtain page content, user will decide whether a webpage is a login page or bot-detect page since there we are not capable to construct a general page distinguisher. We will let user define their own distinguisher.(This also allow a webpage to be both login page and bot-detect page)

For login page, user will enter login information including the information of login field and corresponding values. Then in loginPageHandler class autoLogin() will send the form contain login

information and jump to the response page. In ContentProcessor class, Advice message view has been added to the getHtmlElement method for autoLogin before getting HtmlElement if the webpage is a login page.

Since there are many different types of login page and bot detect page, we will let user defined the specific format and method of login information and bot detection handling. So we ues partial class and method for those classes.

Sample App

We designed a webcrawler to gather house rental information from https://www.mcgill.ca/students/housing/.
We only use basic page getter feature and text processing feature in this app.