Home task HT3 variant 2.

Отчёт о проделанной работе.

Задание HT3 v2 функционально не завершено по причине неоднозначности выбора технологии или библиотеки для считывания HTML кода и преобразование его (если понадобиться) в строку.

Мною была проделана следующая работа по поиску и пробному применению разных подходов в процессе решения поставленной задачи, в течение которой я столкнулся с некоторыми проблемами, о которых пойдёт речь ниже.

Тестирование производилось на следующих веб страницах: google.com, mail.ru, tut.by

1. Использования класса URL и InputStreamReader (metod readHtmlWithUrlAndInputStream()):

Проблемы:

* Google.com предоставляет видимый текст страницы на белорусском языке и для корректного отображения текста требуется явная перекодировка в “Windows1251”.
* Страницы mail.ru и tut.by закачиваются только script – ты и некоторые теги, но не вся страница.

2. Использование класса Url и BufferedReader (method readHtmlWithUrlAndBufferedReader())

* Данный способ достаточно хорошо справляется с вопросом закачки HTML страницы, однако для страницы google.com с белорусским текстом требуется явно перекодировать текст в кодировку “Windows-1251”. Для автоматизации данного процесса не хватило времени.

3. Использование Apache HttpClient (method readHtmlWithApacheHttpClient())

* С кодировкой сайта google.com в белорусской локализации справился на отлично
* Однако как и в первом случае возникли проблемы с загрузкой HTML кода со страниц mail.ru и tut.by: закачиваются только script – ты и некоторые теги, но не вся страница.

4. Используя JettyClient (method readHtmlWithJettyClient()) не удалось добиться работоспособности метода.

5. Использование HtmlCleaner (method readHtmlWithHtmlCleaner()):

* На странице goole.com с белорусской локализацией проблемы с кодировкой.
* Страницы tut.by и mail.ru вообще не закачало.

6.Использование библиотеки Jsoup (method readHtmlWithJsoup())

* Страница google.com в белорусской локализации отображается с нужной кодировкой без дополнительных манипуляций;
* Страницы mail.ru и tut.by закачивает не полноценные страницы.

7. Использование WebClient и HtmlPage из библиотеки HtmlUnitDriver (method readHtmlWithWebClientAndHtmlPage):

* Для google.com с бел. локализацией кодировку подбирает сам нужную и сохраняет корректно код
* Для mail.ru не полная страница и в самом начале закачивает script – ы
* Для страницы tut.by не может получить страницу по Url – и выводит список ошибок.

**После выполненной работы мною было принято решение использовать метод с URL + BufferedReader. Так как он будет работать с большинством web страниц.**

Источники:

<http://zetcode.com/articles/javareadwebpage/>