ENGENHARIA

Data Integration

Licenciatura em Engenharia Informática: 2º ano - 2º semestre 2019/2020

Practical Class nº 4
Implementing Wrappers,

1. Introduction

This worksheet is intended for students to learn how to extract data from the web using the HttpRequest function. In Moodle you will find a file with the implementation of this function:

- **`HttpRequestFunctions.java`**

There are two functions, where the only difference is the encoding used for the characters, which can vary from site to site (UTF-8 or iso-8859-1)

2. Exercises

2.1. Test HttpRequest1 function

Create a new Java project named **Class4**.
In the main class implement this test function

```
public static void class4a(){
        String link = "https://eden.dei.uc.pt/~abs/ID/pessoas.html";
         String search ="";
        HttpRequestFunctions.httpRequest1(link, search, "output.txt");

}
```

Copy the **`HttpRequestFunctions.java`** to the folder **src/class4** of the netbeans project
Call function in main
See if the **output.txt** file was correctly created with the source of the website

2.2. Wrappers: extract data from web dados da web

Copy the file HttpRequestFunctions.java to the folder src/class3 of the Class3 Project developed in previous class. Correct any errors (package name).

Delete the **people.html** file from the project folder. The **people.html** file used in the wrappers of last class is now available on the website: **https://eden.dei.uc.pt/~abs/ID/pessoas.html**

Add to the code to call to the **HttpRequest1** function so that the source of the file is temporarily stored on disk before calling the Wrappers. When writing the source to disk, keep the old file name (people.html)

Test the wrappers and verify that the information is extracted from the Web correctly.

2.3. Wrappers: extract data from web, bookstore Bertrand.pt

In your browser, go to **bertrand.pt** and do a search using "Paul Auster". Analyze the source of the web page and identify where are the names and prices of the books returned by the search.

In your Class4 project created at the beginning of the class, create a new Java Class called **Book** with two fields

- title - String
- price - double

Create the constructor, getters and setters.

a) Implement the function:
```
public static ArrayList<Book> searchBooks(String author)
```

- Implement these steps:

- Call the function `httpRequest1` to obtain the source of the site:
  “https://www.bertrand.pt/pesquisa/ “ the author to search is the argument of the function

  In this function create a file named **Bertrand.html** with the *source* of the requested webpage

- Analyze the file **Bertrand.html** and build the regular expression able to extract the title and price of each book.

- Complete the function:
  - Compile the RE using Class Pattern;
  - Creatre a Scanner/FileInputStream to read the file **Bertrand.html**;
  - Read one line at each time and use class Matcher to search the RE in that line.
  - Extract the title and price using groups.
  - Use the constructor to create na instance of the Book class
  - Add that book to an ArrayList;
  - At the end close the file and return the ArrayList;

- In the main function call the ***searchBooks*** function and print the result. Check if the 20 books have been found.

```
String author = “Paul Auster”;

ArrayList<Book> res = new ArrayList();
res = searchBooks(pesquisa);

System.out.println(“Listing of “ + res.size() + “ found books”);
for(int i=0; i < res.size(); i++){
      System.out.println(“Title : “ + res.get(i).getTitle());
      System.out.println(“Price : “ + res.get(i).getPrice() + “\n”);
}
```

○ Test with other writers: José Luís Peixoto, Fernando Pessoa, J. K. Rowlling, Alice Vieira, etc.