CISC 3115 Project-1

Broken Link Detector

A broken link is an url that references a non-existing document. Broken links are an indicator of poor quality of web pages. Modify the web crawler in the textbook (shown below) such that it identifies and outputs broken links on the website of a given URL. A link can have a full name, as in `https://www.cnn.com/index.html`, or a short name that is relative to the hosting page, as in `<ahref="project1.pdf">`.

```
import java.util.Scanner;
import java.util.ArrayList;

public class WebCrawler {
  public static void main(String[] args) {
    Scanner input = new Scanner(System.in);
    System.out.print("Enter a URL: ");
    String url = input.nextLine();
    crawler(url); // Traverse the Web from the a starting url
  }

  public static void crawler(String startingURL) {
    ArrayList<String> listOfPendingURLs = new ArrayList<>();
    ArrayList<String> listOfTraversedURLs = new ArrayList<>();

    listOfPendingURLs.add(startingURL);
    while (!listOfPendingURLs.isEmpty() &&
        listOfTraversedURLs.size() <= 100) {
      String urlString = listOfPendingURLs.remove(0);
      listOfTraversedURLs.add(urlString);
      System.out.println("Crawl " + urlString);

      for (String s: getSubURLs(urlString)) {
        if (!listOfTraversedURLs.contains(s) &&
                !listOfPendingURLs.contains(s))
          listOfPendingURLs.add(s);
      }
    }
  }

  public static ArrayList<String> getSubURLs(String urlString) {
    ArrayList<String> list = new ArrayList<>();

    try {
      java.net.URL url = new java.net.URL(urlString);
      Scanner input = new Scanner(url.openStream());
      int current = 0;
      while (input.hasNext()) {
        String line = input.nextLine();
        current = line.indexOf("http:", current);
        while (current > 0) {
          int endIndex = line.indexOf("\"", current);
          if (endIndex > 0) { // Ensure that a correct URL is found
            list.add(line.substring(current, endIndex));
            current = line.indexOf("http:", endIndex);
          }
          else
            current = -1;
        }
      }
    }
    catch (Exception ex) {
      System.out.println("Error: " + ex.getMessage());
    }

    return list;
  }
}
```