# Bug report





WebPageExtraction function in ExtractLinkModule

Status: COMPLETED Reported: Oct 28 by Daisy

Product: SearchEngine Component: preprocess module

Version: 1.0

Platform: PC Windows 10

Importance: MAJOR Assigned to: Matthew

How to reproduce:

1. Insert “System.out.println(link\_str);” before return the string of “link\_str”.
2. Run the program
3. Input “Y” after the console display “Do you add new webpages?”.

Bug description:

It displays some invalid links.

For example: msdnaa.cs.cityu.edu.hk/protected/openidc?iss=https%3A%2F%2Flogin.microsoftonline.com%2F83eef0ff-65ef-4870-89dc-e1b9c8e4356f%2Fv2.0&target\_link\_uri=https%3A%2F%2Fmsdnaa.cs.cityu.edu.hk%2Flogin%2F

Possible Solution:

In the ExtractLinkModule class, the function “WebPageExtraction” does not check all the possible invalid hyperlinks. The hyperlink should only contain one full stop and should not contain strange signs. More conditions should be considered in the function “WebPageExtraction”.