ST0316 ADVANCED JAVA PROGRAMMING
ASSIGNMENT PART 2 (25%)
Mini Web Crawler

1. The Problem

In assignment 2, you will code and implement a simplified web crawler which will help you retrieve web documents from the Internet based on a search phrase.

From Wikipedia: A Web crawler starts with a list of URLs to visit, called the *seeds*. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the *crawl frontier*. URLs from the frontier are recursively visited according to a set of policies.

Your simplified web crawler application is required to provide the following functionalities:

1) Enable user to enter a search phrase which may **contain more than** one word.
2) Send the query to two search engines of your choice **concurrently** such as yahoo or bing.
3) Program will need to analyze the downloaded html source from the first 2 search results from each search engine. This will form the **seeds**.
4) Program will need to analyze the downloaded html source from the seeds to find out the top 10 unique webpages through multi-threading.You would have to look for the specific patterns to retrieve the web site address **eg http://....** as a pattern. You will need to apply suitable regular expressions to find such patterns.
5) Ignore advertisements when processing the web content.
6) Your application will create **2 separate threads** which will **download** and process each web document to find the web URLs, which will be added to the Queue.
7) Each of these 10 webpages URLs will be added to an appropriate data structure such as Queue as they are found.
8) The 2 threads will process the next available and unprocessed web URL once it has finished its current task. They will keep doing so, until the 12 websites are found with its contents downloaded.
9) The 10 website URLs and its html page contents **(saved locally**) should be shown to the user through a GUI (For example, after clicking on a selected URL, the web page content will be shown in a text area). The website URLs are to be displayed in a list in ascending order.
10) Keep track of the number of occurrences of the search phrase within the html page and display the number of occurrences.

Note:
You are given PageRead.java. A static method PageRead(..) is provided which helps you download the webpage and return the content (page source) as a StringBuilder. For example if you call PageRead("http://www.cnet.com"); in your java code, this method will return you a StringBuilder representing the html source page data from that webpage.

2. Assessment

This assignment constitutes 25 % of the entire assessment for this module. You will be assessed as a group as well as an individual based on the following:

- Group Work
- Independent Research/Study
- Peer Evaluation
- Contribution
- Graphical User Interface
- Use of Multi-threading
- Processing and storing of relevant data
- Proper usage of collection framework classes like Queue
- Use of regular expressions
- Successfully displaying the results to the user
- Other features/functions

Submission Details
==========
*Deadline: 7th **Aug 2017, 08:30am***
Submit through: Blackboard

Interview Details
==========
*Week 17, 7th **Aug 2017 to 11 Aug 2017***

Late Submission
==========
50% of the marks will be deducted for assignments that are received within ONE (1) calendar day after the submission deadline. No marks will be given thereafter. Exceptions to this policy will be given to students with valid LOA on medical or compassionate grounds. Students in such cases will need to inform the lecturer as soon as reasonably possible.
Students are not to assume on their own that their deadline has been extended.

Warning: Plagiarism means passing off as one's own the ideas, works, writings, etc., which belong to another person. In accordance with this definition, you are committing plagiarism if you copy the work of another person and turning it in as your own, even if you would have the permission of that person. Plagiarism is a serious offence and disciplinary action will be taken against you. If you are guilty of plagiarism, you may fail all modules in the semester, or even be liable for expulsion.

You are encouraged to provide extra advanced features for your application. Examples would be:

1) Allow user to specify the number of threads to download and process the web page source.
2) Compare performance of using say 1 vs 2 vs 4 threads, by keeping track of the total time to process requests.