1. **Web Crawler Implementation**

The application should be able to:

* Download a given webpage from a website
* Using appropriate techniques, extract the hyperlinks found on the downloaded page
* Store the links in a database
* Fetch new links from the database and display in a UI
* Continue to crawl the new links found

Notes:

* Use multithreading and event handling where it is feasible
* The application must compile and run in Visual Studio 2010 or 2012 (must include the data store added to the project, as well as all necessary libraries and resources)
* As a guideline, you should spend maximum 8 hours in total to develop the application

1. **Code Review**

**Code contains following modules:**

1. Downloader
   1. This module will download the web page and extract the links from the page using the HtmlAgilityPack DLL.
2. Crawl WebPage
   1. This module has information about crawl page
3. Components
   1. Multithreaded component handles the multiple threads to handle the crawling
   2. Queue component will feed the links to the multiple threads for crawling
4. Database
   1. Create one table called ‘crawl’ you can find the create script at ‘Pigo\Database\Create\_Table.sql’

Note: Please change the app settings as below:

**<appSettings>**

**<add key="MaximumThreads" value="10"/>**

**<add key="ConnectionString" value="Data Source=localhost;Initial Catalog=Crawler;User ID=sa;Password=Password!"/>**

**</appSettings>**

**Change ConnectionString according to SQL server setup.**

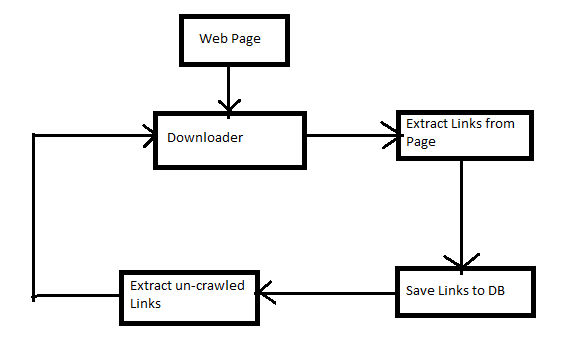
**Weakness:**

1. Validations are not properly handled in the code. i.e. validation about web page content and crawling validations
2. Store Procedures are not used, Indexing for table is not done
3. HTML Parser is not written by own.

**Improvement Areas:**

1. Write own efficient HTML Parser
2. Crawl high rank web pages ahead of normal pages
3. Write an Algorithm for Re-Visit crawling Policy
4. Write Reinforcement Machine learning algorithm for focused crawling using some pre training data.
5. Write URL caching techniques for web crawling.
6. **Process Notes**

**Crawling strategy:**



Downloader is implemented using Multithreading and Queue techniques. To extract links from the given Page I have used ‘HtmlAgilityPack’ DLL (‘*Pigo\Library\HtmlAgilityPack.dll’*)

For saving Links to database I have used SQL Server 2008. Create Database with the name of ‘Crawler’ and use Create table script from the ‘Pigo\Database\Create\_Table.sql’

I have put check in the application so that it can’t crawl the same page again.

GUI will continuously display the links that are queued up for crawling.