Information Retrieval Project

## **Module 1: Web Crawler**

### **Delivery Time:**

# Crawler Requirements

1. The crawler should download **at least 3000** documents from the website: [**www.cnn.com**](http://www.cnn.com) or [**www.bbc.com**](http://www.bbc.com) or[**www.wikipedia.com**](http://www.wikipedia.com)
2. You can use any of the following to build **store the page content**:

* SQL Server 2008/2010/2012

1. ONLY **English Web pages** are stored. Ignore pages with other languages

// search the web for function that detect the language

1. C# or Java might be used
2. Multithreading is BONUS

// your crawler starts with more than one seed

# Crawling Processing Steps

1. Put a set of known sites on a queue (Seeds).
2. Fetch the document at the URL.
   * Sample Code for **fetching** the page (C#)

Using System.Net;

String URL=“http://www.cnn.com”;

// Create a new 'WebRequest' object to the mentioned URL.

WebRequest myWebRequest = WebRequest.Create(URL);

// The response object of 'WebRequest' is assigned to a WebResponse' variable.

WebResponse myWebResponse = myWebRequest.GetResponse();

Stream streamResponse = myWebResponse.GetResponseStream(); StreamReader sReader = new StreamReader(streamResponse);

string rString = SReader.ReadToEnd();

streamResponse.Close();

sReader.Close();

myWebResponse.Close();

1. Parse the URL – HTML parser
2. Extract links from it to other docs (URLs)
   * Sample code for **parsing** the page and **extracting** link from it

// at first add reference to mshtml from solution explorer

using mshtml;//download it if it is not available

IHTMLDocument2 myDoc = new HTMLDocumentClass();

myDoc.write(rString);

IHTMLElementCollection elements = myDoc.links;

foreach (IHTMLElement el in elements)

{

string link = (string)el.getAttribute("href", 0);

}

1. For each extracted URL
   * Ensure it passes certain URL filter tests
   * Check if it is already exists (duplicate   
      URL elimination)
2. **NOTE: Please consider lab notes conducted by TAs and refer to (section2.ppt) for more details**

**Good LUCK J J J**