CS 2026 – Spring 2009
Assignment #1
1/23/2009
Due: Friday 1/30/2009 11:59 PM

In this assignment you are asked to create a C# program to parse out hypertext links in an HTML document.

The HTML source code will be passed to your application as a program argument. For example:

```
:~> MyApplication.exe “<html><body>Go to <a href='http://www.google.com/'>Google</
a></body></html>”
```

Hyperlinks

Hyperlinks are identified in HTML code by any of these patterns:

- `<a href='[URL]'>text</a>`
- `<a href=”[URL]”>text</a>`

Your job will be to look for parts of the passed-in string that match any of the patterns above, and count the number of times a particular website is linked to. To do this counting, you can store the URLs you find in a HashTable or a two-column rectangular array where the first column contains the URL, and the second column contains the number of matches.

When counting, ignore the transfer protocol being used (i.e. the `http://`). You are asked to group the URLs by domain and sub-domain. Disregard specific pages or folders that come after the domain. For example:

- http://www.foobar.mywebsite.com/folder1/folder2/page.php?q=hello should be parsed as foobar.mywebsite.com
- www.google.com and google.com should be both counted as a reference to the same site.
- However video.google.com should be counted as a different site than google.com or www.google.com

Your program should print to the screen a listing of each domain and the number of times it was linked to in the text. An example of an output would be:

google.com	2
video.google.com	1
cornell.edu	3
cs.cornell.edu	5

Assumptions You Can Make

You can make the following assumptions:

- The input to your program will be a well-formatted HTML source code.
- The input to your program will all be in lower-case characters.

Tips

Here are some tips you might find helpful in completing this assignment:

- The following MSDN article has some helpful tips on how to search strings in C#: http://msdn.microsoft.com/en-us/library/ms228630(VS.80).aspx
- The following MSDN article has many links to helpful string-related articles: http://msdn.microsoft.com/en-us/library/ms228364(VS.80).aspx
- The following MSDN article has some general information about C# arrays: http://msdn.microsoft.com/en-us/library/9b9dty7d(VS.80).aspx
- If you use a two-column array to store your parsed results, you should think of boxing/unboxing as discussed in Friday's lecture specially that you will need to store a string URL and an integer counter.
- When you parse out a URL from the input text, use some of the string methods in the articles above to get rid of the "`http://`" and whatever comes after the first "/" in the URL (this will just leave you with the top-level URL of a page.
- Since we are counting "www.google.com" and "google.com" both as links to the same URL, then it will be very helpful for you to remove the "`www.`" from the URL before putting it in your data structure, or incrementing its count.
- If you need any help, talk to me after class on Monday or come to my office hour on Wednesday.

Why is this useful?

In addition to exercising your C# skills, this assignment will give you first hand experience in doing basic web page analysis. Such, and more complicated, analysis is used continuously by researchers, and commercial companies like Google, Yahoo, and Microsoft to run their search engines and even enhance their advertisement networks.

Academic Integrity Reminder

Remember that you may have general discussions about how to approach this problem with your peers, but you should work on the final solution by yourself alone. If you are stuck or are having trouble, you may email me or talk to me after class on Monday or during my office hour on Wednesday.

Good Luck!