WebSPHINX Lab

Aim: Learning to customize a crawler by working with the interactive WebSPHINX workbench.

WebSPHINX: A personal, customizable web crawler: http://www.cs.cmu.edu/~rcm/websphinx/

Javadoc: http://www.cs.cmu.edu/~rcm/websphinx/#doc,
http://www.cs.washington.edu/education/courses/cse454/02au/web_api/websphinx/Crawler.html

Additional documentation:
http://seanfoy.mikejorgensen.com/truman/summer-research/software/websphinx-0.4a/websphinx-doc/websphinx.Crawler.html

Research Paper: http://www.cs.cmu.edu/~rcm/papers/www7/

Last release was 2002 so not the most robust code for modern websites.

- Exercise 1

Get all the faculty images from Villanova's computer science department.

Examine the structure of our department website to figure out where the faculty page is.

Crawl the subtree http://csc.villanova.edu/faculty and extract <img> to html file ./facutly-images on Pages whose URL matches the expression *faculty*

See faculty-images or the file you just created. You will see that along with the faculty, we also have lots of affiliation logos.

Is there someway we can clean it up?

Let us look at the source of a faculty page and observe the valid faculty images and the logos and see a pattern in their URLs. We notice that the valid faculty images have the pattern http://images.csc.villanova.edu/img/*.jpg. So lets run another crawl with just this change in the extracted field <img src="http://images.csc.villanova.edu/img/*.jpg">

- Exercise 2

Get all the news articles' titles from Villanova computer science department.

Figure out where the news articles are and what html tag the tiles have.

Crawl: http://csc.villanova.edu/news/year/2011

Extract: <h3></h3>

On Pages with URL: *news/view*

The resulting file will be empty. I have no idea why this happened. Either our crawler has buggy code or our cms is doing weird stuff.

Let us try crawling starting with http://csc.villanova.edu/news/ hoping that crawling this bigger sub-tree we would get all the news for all the years. If you examine your file or news-second-try, you will notice we will just get the latest news. This may happen because we have an automatic redirect from http://csc.villanova.edu/news/ -> http://csc.villanova.edu/news/year/2011 and the crawler is crawling the page it is being redirected to.

The only way we can fix this is if we crawl starting from http://csc.villanova.edu and select Crawl: The Server instead of the Sub-Tree. This gets us the result we want. I don't know why we need to say crawl the server instead of the sub-tree even when we start from that high up in the tree. But that is the only way it works.

See news-titles or the file you just created. Looking at the file we also come to know that the total number of news articles on our site is 269.

In any case, we can see how the interactivity of the WebSPHINX workbench can help us see what is going wrong and fix it.

- Exercise 3

Extract all the images from Villanova's gallery section.

See gallery-images

- Exercise 4

Extract the email of all full time faculty.

See faculty-email

- Exercise 5

Crawl a site of your choosing and extract some interesting information.

- Solutions

Exercise 3

Crawl: http://csc.villanova.edu/galleries

Extract: <img src="http://images.csc.villanova.edu/img/reflections/*.jpg">

on URL: http://csc.villanova.edu/galleries/list/*

Exercise 4:

Crawl: http://csc.villanova.edu/faculty/fullTime