**An Insight into Focused Web-Crawler**

A **focused crawler** is a [web crawler](https://en.wikipedia.org/wiki/Web_crawler) that collects Web pages that satisfy some specific property, by carefully prioritizing the [crawl frontier](http://nlp.stanford.edu/IR-book/html/htmledition/the-url-frontier-1.html) and managing the hyperlink exploration process.[[1]](#endnote-1)

In this project, the specific property is given by the keywords which user takes as the input. The user might give one or more keywords and the crawler scrapes the website based on the keyword which might be found in the anchor text or anchor href on the page.

In doing so, extending BFS strategy to focused crawler yields better results. The program must support any variation of the keyword. In other words, the search must be case insensitive search with any prefix or suffix. This task is achieved by analyzing all the anchor tags in a page and using python regular expression to find the string in the page. Once found, those URLs are added to the neighbor's list of the URL crawled.

for key in keyword\_list:  
 for content in data:  
 if re.search(key, str(content), re.IGNORECASE):  
 scraped\_urls.add(href)  
 if re.search(key, href, re.IGNORECASE):  
 if href not in scraped\_urls:  
 scraped\_urls.add(href)

The above piece of code when added to the BFS algorithm, achieves our task of focused crawling.

In addition to that, to support the decompounding strategy, we are using the regex to check the keyword against the anchor tag’s text.

1. https://en.wikipedia.org/wiki/Focused\_crawler [↑](#endnote-ref-1)