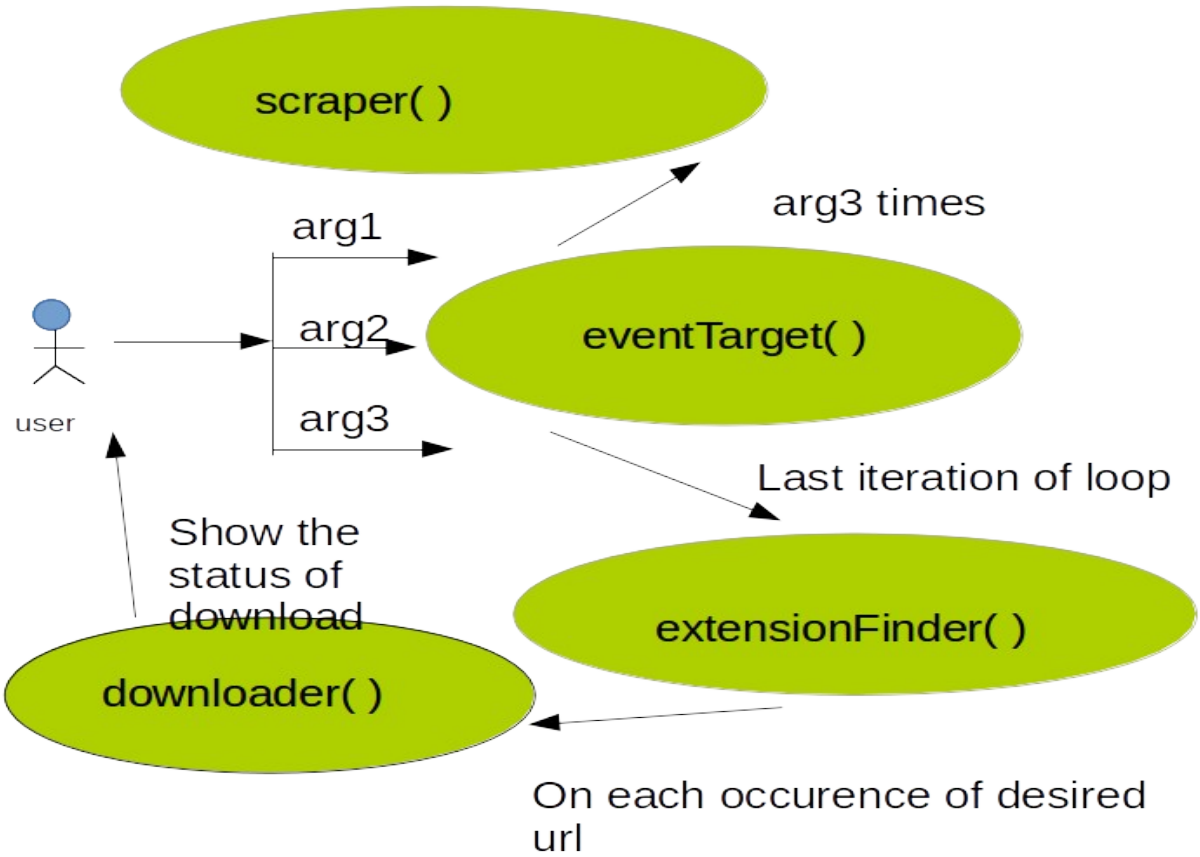
Web Spider to exract data of hidden links of page

Project Description:

The purpose of this project is to help my Data Hero by giving him data of website which is hidden in the links of html document. Like in the pagination division of the any webpage, contents of the division change but problem is that link of the page is not changes because the content are coming dynamically.

Use Cases and Edge Conditions:

To solve the given problem until user want, I will be keep on submitting the form which is called when click the peginated index of dynamic web pages. After submitting the form I get new html content, pass it to a function which will be parsing this html content and extract all the href urls except those who are starting with `javascript'. And then I will be storing these urls in the list so that I can keep track, which all links I will have to visit without repeating the link. Then on the last iteration of the loop I will be calling another function which will take urls of that list and one by one parse their html content and extract those links which are having the required extension and call the another downloader function on this link to download the files.



UML Diagram of the solution

There are some javascript function calls in the href of <a> tags of dynamic pages and I will have to filter these otherwise they will be cause of error because they are not following normal html href pattern.And other I will have to follow the convention to convert the relative urls in to absolute urls so that no repeatition of urls occurs. If there are no links of perticuler type of files in the html document then it will show some kind of message. And there will be a case of multiple occurence of the same url of file then I will have to handle this situation.

Workflow:

I tried to give maximum flexibility to user.To run this script user will have to give three arguments at runtime,first is the website url in which the hidden link exist, second is On how many pages he/she want to get the data and last is what kind of file he/she want to download (.pdf, .csv or .mp3).
And in the output from, script will show the found desired link to user and starting downloading with real time status of file such as size of file, how much is downloaded in byte and in percent.
In order to solve this problem I will use the spider algorithm of web crawling. I will use some external libraries of python such as parsing libraries i.e. lxml.etree or Beautiful and automated javascript form submition library i.e. mechanize and some in-built libraries of python such as urllib/urllib2, urlparse, time, re etc. In the eventTarget function I will use the mechanize, re, time libraries.With the reguler expression I will extract the parameter sent to form for setting __EVENTTARGET.value from javascript hrefs and with the help of mechanize I will assign them directly and submit the form by disable all the javascript controls. This process will continue count entered by user and each iteration I will send this html content to scrapper function. In the scrapper function I will parse the that html content with the help of BeautifulSoup library and scrap all the href of divison whose id is 'news_content_mid'.
Now will put these scraped hrefs in to a list to ressolve the duplication issue for any generic page. And In the last iteration of main loop of the eventTarget function I will send list which contains the hrefs to a function extensionFinder function. Now in this function a loop will iterate on the list and in each iteration I am using urllib/urllib2 to get the source code of the urls of the list and again with the help of BeautifulSoup library I will get that hrefs only which are having the desired extension in their path part of href that I am checking with the help of urlsplit function of urlparse library. If the extension matches then send this href to downloader function which will extract the filename and calculate the size of the file and provides the real time status of the downloading, I will use the urllib2 library to download the files.

Data formats and Reporting:

In the output I will show the filename and size of the file which is downloading and the current status of ongoing download. Such as how much is byte are downloaded and how much percentage of file is downloaded. And after each download, in the output total number of files downloaded till that files.

Performence and scaling:

Since in start I will collect those links which can contain the desired links(i.e. '.pdf' links).So initially it will take time but when downloading starts it will be very fast definetly. It would be better if user will have a good internet connection and good processer in machine. It wil be a platform independent can run on any machine having required external library.

Unresolved issues:

I want to make it completely generic but then it will dwonload those files which are out of peginated division because div name could be anything dependent on the developer of that web page. And the issue is that text of the next button could be anything dependent on webpage. So if somehow we can take these two value from the user then it will be completely platform independent as well as webpage independent.