User Manual

Requirements:

Since it is a operating system independent crawler so you can use it on any os but the required dependencies are -

i)Python2.7

To install the python2.7 on your operating system, you can follow these given links.

Windows:

To check if you already installed the python2.7, enter the python in command promt then if you don't find the python interpreter on cmd promt then follow these links.

https://wiki.python.org/moin/BeginnersGuide/Download
http://docs.python-guide.org/en/latest/starting/install/win/
https://www.youtube.com/watch?v=gD4eulxGNok

Linux:

Although in the linux python2.7 is installed by default. To check if the python install open the terminal and type the python, then python interpreter will be open with all details about the version of python. But if you don't have then follow these links.

https://wiki.python.org/moin/BeginnersGuide/Download
http://docs.python-guide.org/en/latest/starting/install/linux/
http://www.linuxfromscratch.org/blfs/view/svn/general/python2.html

ii)Beautiful Soup

To install the Beautiful soup you can these links.

http://thyagjs.blogspot.in/2013/07/how-to-install-beautiful-soup-or-bs4-on.html
http://babydatajournalism.tumblr.com/post/20710688328/download-and-install-beautiful-soup

iii)Mechanize

To install machenize on ubuntu you can directly type the command in the terminal-

```
sudo apt-get install python-mechanize
```

To install on windows you can follows these links

http://wwwsearch.sourceforge.net/mechanize/download.html
http://stackoverflow.com/questions/4888463/how-to-install-mechanize-for-python-2

How to run:

Now you have all the dependecies so you will be able to run this crawler nad get the desied data from the hidden links of html documents.

Ubuntu :

open the terminal by pressing the 'ctrl + t' , and then go to the directory where the 'crawler.py' exist. Now type the following command on terminal

```
python crawler.py  <website name>  <no. Of pages >  <extension of files>
```

Windows:

Go to search box and type 'cmd' press enter, then command promt will be open now to run the crawler go to the directory where the 'crawler.py' exist. Now type the following command in cmd promt.

python crawler.py <website name> <no. Of pages > <extension of files>

Example:

Suppose you want to download the content of only five pages of the website http://trai.gov.in/Content/PressReleases.aspx and desired file extension is '.pdf'. Then type this command in the terminal/comman promt

python crawler.py http://trai.gov.in/Content/PressReleases.aspx 5 .pdf

On running time you can see total number of files downloaded till now and the name and size of that file which is downloading with its real time ongoing status with size and percentage. And when the download completed then you will find a directory whose name will be 'PDFs' that contains all the files.

Relevent Features:

1) You will be able to download any type of files by just entering the extension of file

2)You have control the downloading by entering the page number . Suppose you want the data of only first two pages then just enter the 2 in place of no.of pages.

3)After downloading your files won't be scatterd in whole directory. They will be downloaded in a seprate folder in the same directory, name of that folder changes on the basis of extension of files.

4)At the running time you can see total number of files downloaded till now and the name and size (in bytes) of that file which is downloading with its real time ongoing status with size and percentage.

5)If you are unable to download the files because the 'proxyAuthenticated network' then open the file proxy.py in any text editor and enter your 'username' , 'password' , 'host/ipaddress' and ' port' in the fields provided. And now open the file crawler.py in the editor and uncommnet(remove the #) the text '#from proxy import urllib2 'and comment(add the # before import) the 'import urllib2'. Now after change these settings, you are using a proxy opener . Now you will be able to download the files.