Multithreaded Web spider in Python

Group 4 Members

Wepngong Ngeh Benaiah	**FE12A191**
Esambe Elvis Njume	**FE12A056**
Nkeangnyi Tonia	**FE12A138**
Kingue Patrick	**FE12A087**
Takoungang Dieudonne	**FE12A172**
Naoussi Martial	**FE12A114**

Requirements:

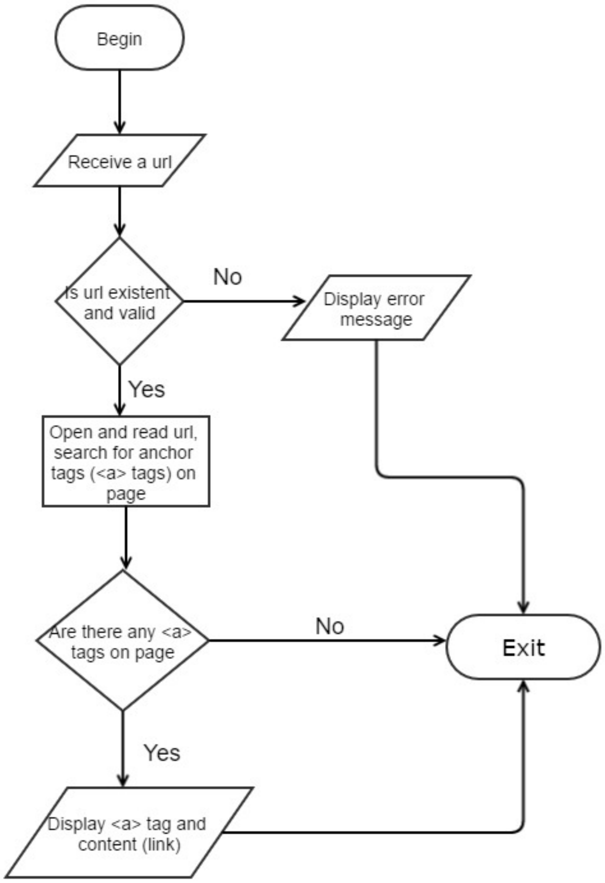
Hardware

- RAM: 4GB
- Hard drive: 500GB
- Processor: 2.2GHz

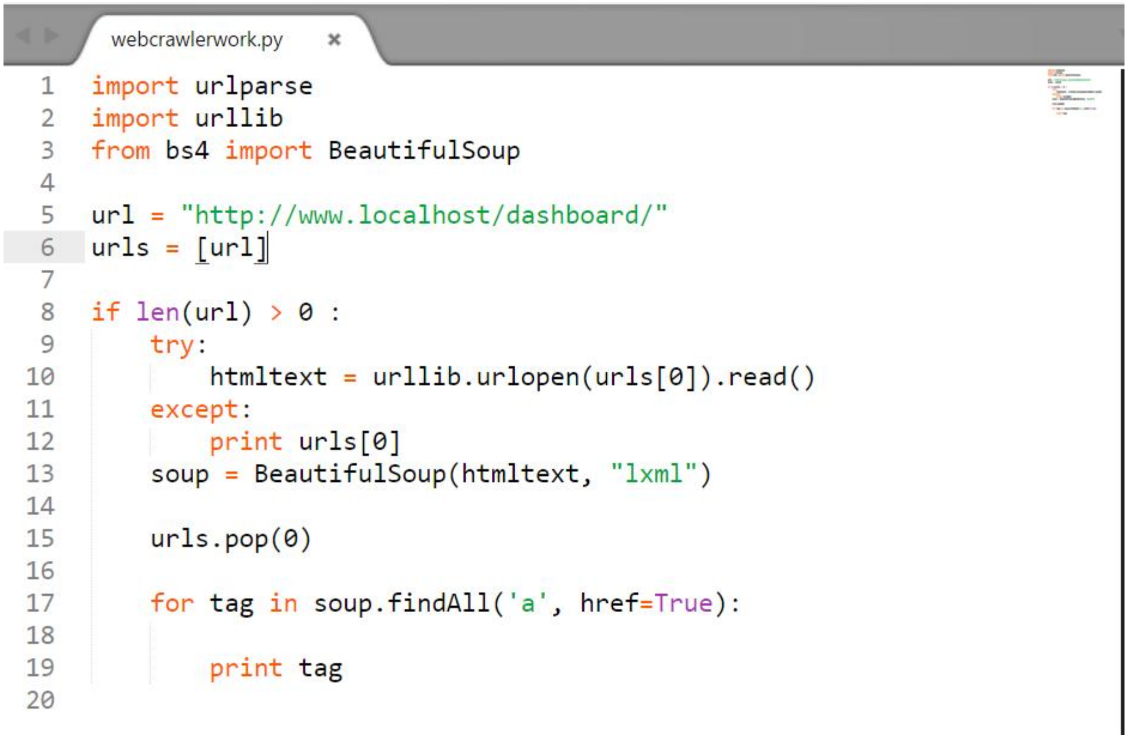
Software

- Operating System: Windows 10 Home and Kali Linux Rolling
- Text Editor: Notepad++/sublime text
- Language: Python 2.7
- Documentation: Microsoft Office Word
- ***Libraries Needed:***
  - Urlparse
  - Urllib
  - Beautiful Soup

Flowchart

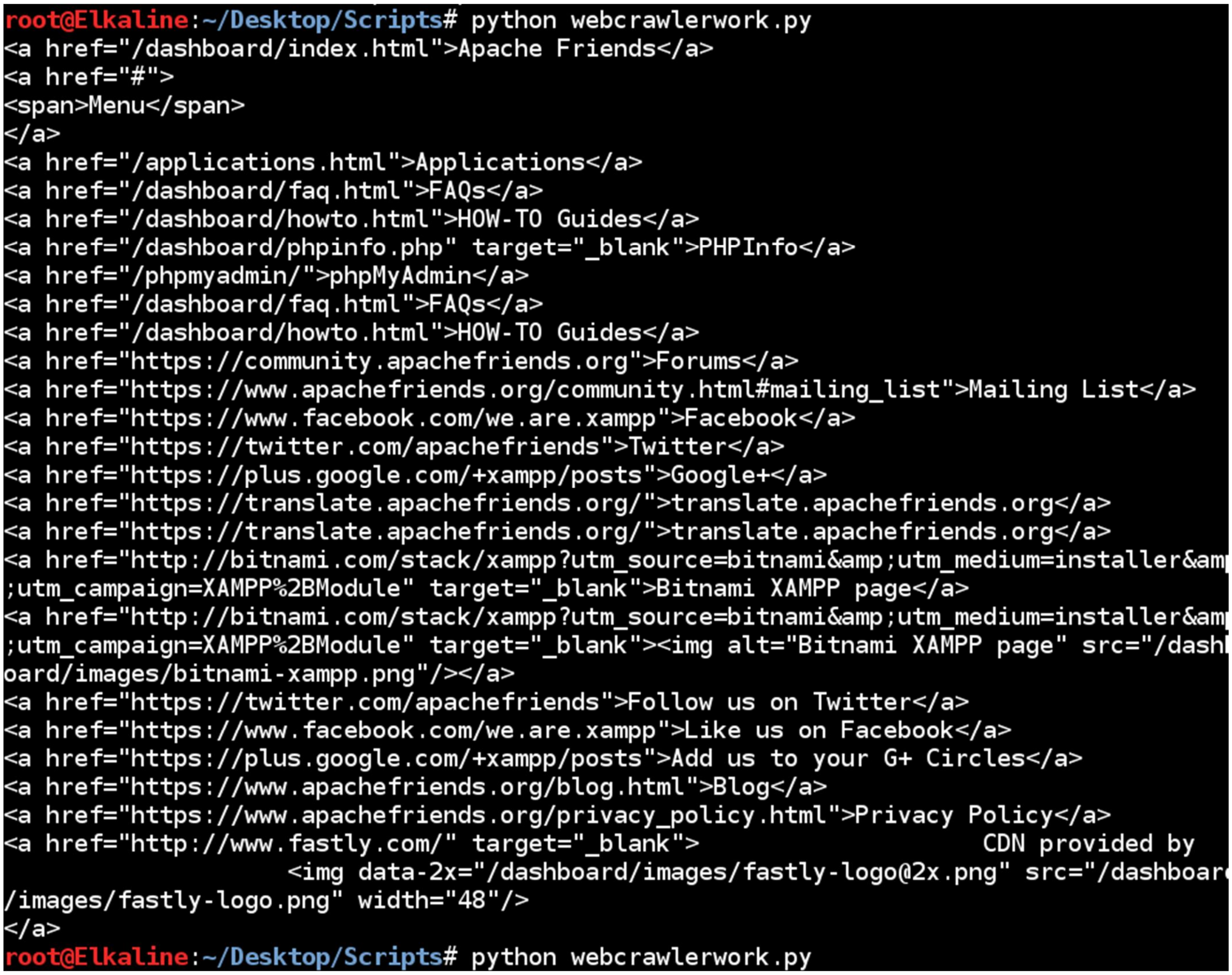

Begin
Receive a url
Is url existent and valid
No
Display error message
Yes
Open and read url, search for anchor tags (<a> tags) on page
Are there any <a> tags on page
No
Exit
Yes
Display <a> tag and content (link)

Source Code Snippet



```
import urlparse
import urllib
from bs4 import BeautifulSoup

url = "http://www.localhost/dashboard/"
urls = [url]

if len(url) > 0 :
    try:
        htmltext = urllib.urlopen(urls[0]).read()
    except:
        print urls[0]
    soup = BeautifulSoup(htmltext, "lxml")

    urls.pop(0)

    for tag in soup.findAll('a', href=True):

        print tag

```

Output



```
root@Elkaline:~/Desktop/Scripts# python webcrawlerwork.py
<a href="/dashboard/index.html">Apache Friends</a>
<a href="#">
<span>Menu</span>
</a>
<a href="/applications.html">Applications</a>
<a href="/dashboard/faq.html">FAQs</a>
<a href="/dashboard/howto.html">HOW-TO Guides</a>
<a href="/dashboard/phpinfo.php" target="_blank">PHPInfo</a>
<a href="/phpmyadmin/">phpMyAdmin</a>
<a href="/dashboard/faq.html">FAQs</a>
<a href="/dashboard/howto.html">HOW-TO Guides</a>
<a href="https://community.apachefriends.org">Forums</a>
<a href="https://www.apachefriends.org/community.html#mailing_list">Mailing List</a>
<a href="https://www.facebook.com/we.are.xampp">Facebook</a>
<a href="https://twitter.com/apachefriends">Twitter</a>
<a href="https://plus.google.com/+xampp/posts">Google+</a>
<a href="https://translate.apachefriends.org/">translate.apachefriends.org</a>
<a href="https://translate.apachefriends.org/">translate.apachefriends.org</a>
<a href="http://bitnami.com/stack/xampp?utm_source=bitnami&amp;utm_medium=installer&am
;utm_campaign=XAMPP%2BModule" target="_blank">Bitnami XAMPP page</a>
<a href="http://bitnami.com/stack/xampp?utm_source=bitnami&amp;utm_medium=installer&am
;utm_campaign=XAMPP%2BModule" target="_blank"><img alt="Bitnami XAMPP page" src="/dash
oard/images/bitnami-xampp.png"/></a>
<a href="https://twitter.com/apachefriends">Follow us on Twitter</a>
<a href="https://www.facebook.com/we.are.xampp">Like us on Facebook</a>
<a href="https://plus.google.com/+xampp/posts">Add us to your G+ Circles</a>
<a href="https://www.apachefriends.org/blog.html">Blog</a>
<a href="https://www.apachefriends.org/privacy_policy.html">Privacy Policy</a>
<a href="http://www.fastly.com/" target="_blank">                    CDN provided by
                    <img data-2x="/dashboard/images/fastly-logo@2x.png" src="/dashboar
/images/fastly-logo.png" width="48"/>
</a>
root@Elkaline:~/Desktop/Scripts# python webcrawlerwork.py
```