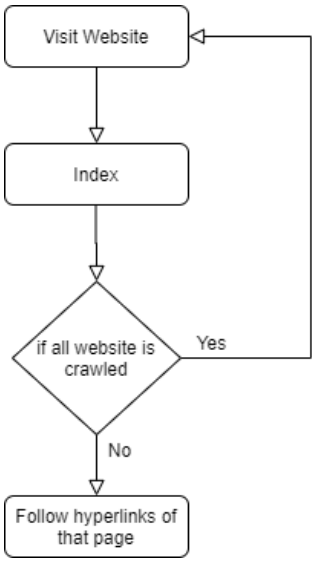
Crawler Design

Author: Jibreel Natsheh – Intern at ASAL Technologies and 2nd year Computer Engineering student at Palestine Polytechnic University

Date: 7/7/2020

What is the crawler supposed to do:

1. Visit Certain Website
2. Index that page
3. Follow all other links in that page
4. Re-Perform 2 and 3 for all other links in that page
5. Visit other Website



Methodology:

1. We will use Python Language cause of the large and useful libraries that could help in requesting and dealing with html and web pages.
2. Important Libraries and drivers:
    a. Gecko Driver: is a driver that handles http/ https requests and contact with browsers we use selenium.webdriver library to deal with it in python.
    b. Use Firefox and prefer Firefox for Tor browser to connect with using the geckodriver in order not to able to start all over again in case you got banned from the website
    c. Use Beautiful soup (bs4) library to read the page source from the web driver
3. We may use either BFS or DFS when crawling and visiting the hyperlinks of the website and the main page (we first visit) would be the root of that tree and the hyperlinks in this page is the children to this node and for each children all hyperlinks in that page would be children to it and grandchildren to the root and so on but if a certain hyperlink exists in higher level or sibling level of the tree it won't be added to the tree, this way we grantee that the crawler won't end in an infinite loop, and also this way we make sure that we have visited the whole website.