**Identify**

Problem : Identifying publically searchable links and then labelling them appropriately.

Approach :

Step 1: Find all the links from any of the avaible Searching Engines.

I used Bing.com as it is easy to parse the results and get the links directly.

The user is asked for the search parameter i.e the name and other information which will be used for segeregating the pages.

The python module Beautiful Soup was used for parsing the pages and then extracting all the available links within the search result.

Step 2 : Parsing each of the links that we got from the above step by converting the html to plain text and by removing the code that is html,css,javascript from the page source.

We use a header file for crawling through websites such as linked because they don't allow crawlers to get past.

We then get the source of the page and then clean it accordingly.

Then it is passed over to next step where the factors are found in the page source

Step 3: We get the cleaned page source as a parameter and the link as another parameter. We then search for the factors variables in the page source to find the relevancy.

The matching method is very simple that i have used. It can however be improved in accuracy with the help of NLP. Which i haven't studied yet.

The match method is little inaccurate because in social networking sites we can find your name in the friend list of another person which increases the relevancy but it isn't relevant much.

Sample Run:



