**ADA – A2 – Q5:**

Craig Fraser, 15889604

Connor Hewett, 15903849

Overview:

After running SearchTool.py, you will be asked to enter seed URLs.

Entered seed URLs must be full urls. EG: <http://www.seed.com>

Once you've entered the seed URLs you want, press enter without typing anything in to continue.

You will then be asked to enter a keyword to search for. Do that.

The spider will then begin crawling. It will list each URL it visits as it visits them.

Once it has reached the depth-limit of 3, or run out of links to follow, it will list the sites with keyword matches, ordered by page-rank.

SpiderLeg:

This class contains the methods for scraping data from websites. It has separate methods for scraping links, metadata, images, and titles.

The getHyperLink method also validates any links it finds before returning them. It checks to ensure they are valid urls, and ensures they are not on the exclusion list. The exclusion list includes large, complex sites such as Twitter, Facebook, and Youtube. It also excludes pages in formats that it cannot scrape, such as .asp sites.

Spider:

This class does the actual crawling. It is given a list of seed URLs, and effectively conducts a DFS using the seeds as starting nodes.

It maintains a list of visited and unvisited URLs, as well as an adjacency list.

It checks the keyword metadata of each site it visits to check for matches with the user-selected keyword, and keeps a list of URLs that match.

It also keeps track of the depth of the search, and stops when it hits the limit.

It is also the class responsible for writing the list of visited sites to a txt file.

SearchTool:

This class is responsible for getting input from the user regarding seed URLs and keywords to search for.

It is also the class that calculates the page-ranks.

Finally, it prints the sites that keyword match to the console, and saves a list of them to a text file.