Name: Thanh Tran

Project: Mini web crawler

**Project summary:**

This program is basically a very simplified version of a search engine. Given a seed URL, the program will first analyze it to determine whether the page contains the search query. Then it will “crawl” to all the reference links on this page to continue the search. The searching follow breadth first search algorithm to avoid infinite loop if existed. When the searching reaches a certain level of depth in the network which was predefined by the user, it will end.

Finally, all the results will be sorted based on the number of occurrence of the search query in the web page

**Structure:**

Module Linkfinder.py: the logic of the methods is based on the implementation from <http://www.netinstructions.com/how-to-make-a-web-crawler-in-under-50-lines-of-python-code/>.

This includes how to get all the reference links from an html file of a web page

Module crawler.py: This will create a worker that crawl the network of related web pages following breadth first search

Module page.py: a wrapper class to store the page URL and the count of the occurrence of the search query

Module gui.py: create a GUI to get the user inputs for the seed URL, the search query and the max number of depth in the search process

Instruction: run the gui.py to start the program.

This program includes the concepts that I learned in class. Furthermore, it features networking, data structure and GUI concepts

**Known problems:**

The most obvious problem is the running time. This can be solved by using multi-threading that create more workers that can crawl pages in parallel

I learned some networking concepts

As I mentioned earlier, this project is very simple so I don’t find it difficult

Log file of the program:

Input:

Feed URL: <http://www.wabash.edu>

Search String: College

Max-Depth: 1

**Searching begins**

**Checking http://www.wabash.edu**

**Checking http://www.wabash.edu/news**

**Checking http://www.wabash.edu/calendar**

**Checking http://www.wabash.edu/aboutwabash/directories**

**Checking http://www.wabash.edu/aboutwabash/offices**

**Checking http://www.youtube.com/WabashCollege**

**Checking http://www.facebook.com/pages/Wabash-College/16568430836**

**Checking http://twitter.com/WabashCollege**

**Checking http://www.wabash.edu/news/syndication/rss091feed.xml**

**Checking http://www.instagram.com/wabashcollege**

**Checking https://itunes.apple.com/us/podcast/wabash-on-my-mind/id855321475?mt=2**

**Checking http://www.wabash.edu/mybash**

**Checking http://www.wabash.edu/**

**Checking http://www.wabash.edu/aboutwabash**

**Checking http://www.wabash.edu/aboutwabash/president**

**Checking http://www.wabash.edu/aboutwabash/location**

**Checking http://www.wabash.edu/aboutwabash/history**

**Checking http://www.wabash.edu/aboutwabash/info**

**Checking http://www.wabash.edu/aboutwabash/mission**

**Checking http://www.wabash.edu/aboutwabash/employment**

**Checking http://www.wabash.edu/aboutwabash/wabash**

**Checking http://www.wabash.edu/academics**

**Checking http://www.wabash.edu/academics/majors**

**Checking http://www.wabash.edu/academics/registrar**

**Checking http://www.wabash.edu/academics/pbk**

**Checking http://www.wabash.edu/academics/research**

**Checking http://www.wabash.edu/academics/learning**

**Checking http://www.wabash.edu/plus**

**Checking http://www.wabash.edu/academics/catalog**

**Checking http://www.wabash.edu/academics/library**

**Checking http://www3.wabash.edu/orgs/writingcenter/wc.home.html**

**Checking http://www.wabash.edu/academics/teach**

**Checking http://www.wabash.edu/academics/events**

**Checking http://www.wabash.edu/registrar/transcripts**

**Checking http://www.wabash.edu/admissions**

**Checking http://www.wabash.edu/admissions/after**

**Checking http://www.wabash.edu/admissions/visit**

**Checking http://www.wabash.edu/admissions/requestinfo**

**Checking http://www.wabash.edu/admissions/majors**

**Checking http://www.wabash.edu/admissions/financialaid**

**Checking http://www.wabash.edu/admissions/counselor.cfm**

**Checking http://www.wabash.edu/admissions/apply**

**Checking http://www.wabash.edu/admissions/contact**

**Checking http://www.wabash.edu/alumni**

**Checking http://www.wabash.edu/alumni/voices**

**Checking http://www.wabash.edu/alumni/bigbash**

**Checking http://www.wabash.edu/alumni/wd**

**Checking http://www.wabash.edu/alumni/plannedgiving**

**Checking http://www.wabash.edu/alumni/volunteer**

**Checking http://www.wabash.edu/alumni/alumni\_assoc**

**Checking http://www.wabash.edu/alumni/homecoming**

**Checking http://www.wabash.edu/parents**

**Checking http://www.wabash.edu/alumni/eservices**

**Checking http://www.wabash.edu/alumni/annualgiving**

**Checking http://www.wabash.edu/giving**

**Checking http://www.wabash.edu/giving/annual**

**Checking http://www.wabash.edu/giving/capital**

**Checking http://www.wabash.edu/giving/honorroll**

**Checking http://www.wabash.edu/giving/1832Society**

**Checking http://www.wabash.edu/giving/planned**

**Checking https://www.wabash.edu/giving/give**

**Checking http://www.wabash.edu/news**

**Checking http://www.wabash.edu/calendar**

**Checking http://www.wabash.edu/aboutwabash/directories**

**Checking http://www.wabash.edu/aboutwabash/offices**

**Checking http://www.wabash.edu/admissions/apply**

**Checking http://www.wabash.edu/admissions/visit**

**Checking http://www.wabash.edu/mybash**

**Checking http://www.wabash.edu/news/displaystory.cfm?news\_ID=10771**

**Checking http://www.wabash.edu#featured-carousel**

**Checking http://www.wabash.edu#featured-carousel**

**Checking http://www.wabash.edu/bachelor**

**Checking http://www.wabash.edu/bookstore**

**Checking http://www.wabash.edu/businessoffice**

**Checking http://www.wabash.edu/careers**

**Checking http://www.wabash.edu/bulletin**

**Checking http://www.wabash.edu/hr/jobs**

**Checking http://www.wabash.edu/academics/teach/**

**Checking http://www.wabash.edu/library**

**Checking http://www.wabash.edu/aboutwabash/president**

**Checking http://www.wabash.edu/registrar**

**Checking http://www.wabash.edu/technology**

**Checking http://www.wabash.edu/magazine**

**Checking http://www.wabash.edu/studentlife/**

**Checking http://www.wabash.edu/parents**

**Checking http://www.wabash.edu/visitors**

**Checking http://www.wabash.edu/centers/**

**Checking http://www.wabash.edu/careers/**

**Checking http://www.wabash.edu/admissions/apply**

**Checking http://www.wabash.edu/plus**

**Checking http://www.wabash.edu/news/displaystory.cfm?news\_ID=10571**

**Checking http://www.wabash.edu/news/displaystory.cfm?news\_ID=10821**

**Checking http://www.wabash.edu/news/displaystory.cfm?news\_ID=10892**

**Checking http://www.wabash.edu/news/displaystory.cfm?news\_ID=10892**

**Checking http://www.wabash.edu/news/displaystory.cfm?news\_ID=10872**

**Checking http://www.wabash.edu/news/displaystory.cfm?news\_ID=10893**

**Checking http://www.wabash.edu/news**

**Checking http://www.wabash.edu/news**

**Checking http://www.wabash.edu/calendar**

**Checking http://www.wabash.edu/blog**

**Checking http://www.youtube.com/WabashCollege**

**Checking http://www.wabash.edu/photo\_album**

**Checking http://whatkindofman.wabash.edu**

**Checking http://www.wabash.edu/aboutwabash/facts**

**Checking http://www.wabash.edu/map/**

**Checking http://www.wabash.edu/aboutwabash/president**

**Checking http://www.wabash.edu/bookstore**

**Checking http://www.wabash.edu/hr**

**Checking http://www.wabash.edu/academics/teach**

**Checking http://www.wabash.edu/centers/**

**Checking http://library.wabash.edu/**

**Checking http://www.wabash.edu/international**

**Checking http://www.wabash.edu/careers**

**Checking http://www.wabash.edu/magazine**

**Checking http://www.wabash.edu/bachelor**

**Checking http://www.wabash.edu/businessoffice/**

**Checking http://www.wabash.edu/technology**

**Checking http://liberalartssuccess.org/**

**Checking http://www.wabash.edu/aboutwabash/info**

**Checking http://www.wabash.edu/discrimination/**

**Checking http://www.wabash.edu/accessibility/**

**Checking http://www.wabash.edu/privacy/**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**The following list contains the resulted web pages along with the occurrence of the search string in them:**

**http://www.facebook.com/pages/Wabash-College/16568430836 491**

**http://www.youtube.com/WabashCollege 329**

**http://www.youtube.com/WabashCollege 329**

**http://twitter.com/WabashCollege 201**

**https://itunes.apple.com/us/podcast/wabash-on-my-mind/id855321475?mt=2 196**

**http://www.wabash.edu/aboutwabash/facts 27**

**http://www.wabash.edu/aboutwabash/president 26**

**http://www.wabash.edu/aboutwabash/president 26**

**http://www.wabash.edu/aboutwabash/president 26**

**http://www.wabash.edu/privacy/ 26**

**http://www.wabash.edu/alumni 20**

**http://www.wabash.edu/alumni/alumni\_assoc 20**

**http://www.wabash.edu/news/displaystory.cfm?news\_ID=10872 19**

**http://www.wabash.edu/discrimination/ 19**

**http://www.wabash.edu/aboutwabash/history 18**

**http://www.wabash.edu/giving 18**

**http://www.wabash.edu/news/displaystory.cfm?news\_ID=10771 18**

**http://www.wabash.edu/accessibility/ 18**

**http://www.wabash.edu/aboutwabash/mission 17**

**http://www.wabash.edu/alumni/voices 17**

**http://www.wabash.edu/alumni/homecoming 17**

**http://www.wabash.edu/businessoffice 17**

**http://www.wabash.edu/hr 17**

**http://www.wabash.edu/businessoffice/ 17**

**http://www.wabash.edu/academics/library 16**

**https://www.wabash.edu/giving/give 16**

**http://www.wabash.edu/bookstore 16**

**http://www.wabash.edu/library 16**

**http://www.wabash.edu/bookstore 16**

**http://library.wabash.edu/ 16**

**http://www.wabash.edu/aboutwabash 15**

**http://www.wabash.edu/academics/pbk 15**

**http://www.wabash.edu/admissions 15**

**http://www.wabash.edu/admissions/requestinfo 15**

**http://www.wabash.edu/alumni/bigbash 15**

**http://www.wabash.edu/parents 15**

**http://www.wabash.edu/parents 15**

**http://www.wabash.edu/academics/registrar 14**

**http://www.wabash.edu/admissions/visit 14**

**http://www.wabash.edu/admissions/apply 14**

**http://www.wabash.edu/admissions/apply 14**

**http://www.wabash.edu/admissions/visit 14**

**http://www.wabash.edu/registrar 14**

**http://www.wabash.edu/admissions/apply 14**

**http://www.wabash.edu/international 14**

**http://www.wabash.edu 13**

**http://www.wabash.edu/ 13**

**http://www.wabash.edu/alumni/annualgiving 13**

**http://www.wabash.edu/giving/annual 13**

**http://www.wabash.edu#featured-carousel 13**

**http://www.wabash.edu#featured-carousel 13**

**http://whatkindofman.wabash.edu 13**

**http://www.wabash.edu/academics/teach 12**

**http://www.wabash.edu/registrar/transcripts 12**

**http://www.wabash.edu/giving/1832Society 12**

**http://www.wabash.edu/academics/teach/ 12**

**http://www.wabash.edu/academics/teach 12**

**http://www.wabash.edu/aboutwabash/location 11**

**http://www.wabash.edu/academics/catalog 11**

**http://www.wabash.edu/admissions/financialaid 11**

**http://www.wabash.edu/alumni/eservices 11**

**http://www.wabash.edu/bulletin 11**

**http://www.wabash.edu/technology 11**

**http://www.wabash.edu/centers/ 11**

**http://www.wabash.edu/news/displaystory.cfm?news\_ID=10571 11**

**http://www.wabash.edu/news/displaystory.cfm?news\_ID=10892 11**

**http://www.wabash.edu/news/displaystory.cfm?news\_ID=10892 11**

**http://www.wabash.edu/blog 11**

**http://www.wabash.edu/centers/ 11**

**http://www.wabash.edu/technology 11**

**http://www.wabash.edu/news 10**

**http://www.wabash.edu/aboutwabash/info 10**

**http://www.wabash.edu/aboutwabash/employment 10**

**http://www.wabash.edu/academics/majors 10**

**http://www.wabash.edu/academics/research 10**

**http://www.wabash.edu/plus 10**

**http://www.wabash.edu/admissions/majors 10**

**http://www.wabash.edu/giving/capital 10**

**http://www.wabash.edu/news 10**

**http://www.wabash.edu/careers 10**

**http://www.wabash.edu/hr/jobs 10**

**http://www.wabash.edu/magazine 10**

**http://www.wabash.edu/careers/ 10**

**http://www.wabash.edu/plus 10**

**http://www.wabash.edu/news 10**

**http://www.wabash.edu/news 10**

**http://www.wabash.edu/photo\_album 10**

**http://www.wabash.edu/careers 10**

**http://www.wabash.edu/magazine 10**

**http://www.wabash.edu/aboutwabash/info 10**

**http://www.wabash.edu/calendar 9**

**http://www.wabash.edu/aboutwabash/directories 9**

**http://www.wabash.edu/aboutwabash/offices 9**

**http://www.wabash.edu/mybash 9**

**http://www.wabash.edu/academics 9**

**http://www3.wabash.edu/orgs/writingcenter/wc.home.html 9**

**http://www.wabash.edu/admissions/counselor.cfm 9**

**http://www.wabash.edu/admissions/contact 9**

**http://www.wabash.edu/alumni/wd 9**

**http://www.wabash.edu/alumni/volunteer 9**

**http://www.wabash.edu/giving/honorroll 9**

**http://www.wabash.edu/calendar 9**

**http://www.wabash.edu/aboutwabash/directories 9**

**http://www.wabash.edu/aboutwabash/offices 9**

**http://www.wabash.edu/mybash 9**

**http://www.wabash.edu/bachelor 9**

**http://www.wabash.edu/studentlife/ 9**

**http://www.wabash.edu/news/displaystory.cfm?news\_ID=10893 9**

**http://www.wabash.edu/calendar 9**

**http://www.wabash.edu/map/ 9**

**http://www.wabash.edu/bachelor 9**

**http://www.wabash.edu/aboutwabash/wabash 8**

**http://www.wabash.edu/academics/learning 8**

**http://www.wabash.edu/academics/events 8**

**http://www.wabash.edu/admissions/after 8**

**http://www.wabash.edu/visitors 8**

**http://www.wabash.edu/news/displaystory.cfm?news\_ID=10821 8**

**http://www.instagram.com/wabashcollege 6**

**http://www.wabash.edu/alumni/plannedgiving 6**

**http://www.wabash.edu/giving/planned 6**

**http://liberalartssuccess.org/ 1**

**[Finished in 164.5s]**