Upper Arlington High School

**IB Computer Science HL Internal Assessment**

**For Joe Sheridan**

**Calvin Kinateder**

**IB Computer Science HL**

**Dr. Kahle**

**10 / 06 / 2018**

**Criterion A: Planning**

The problem I am tackling for my IA is for the ComOps department at AEP. ComOps deals with the buying and selling of power. They needed a way to find news about the price of power and really all things that would interest that department on consumer sites. The goal of my program will be to get a list of webpages that could be potentially relevant to a certain keyword. Google does not do this for them because Google a) does not do the best job of sorting by relevance in this instance, and b) does not write to a database. The client’s name is Joe Sheridan and he is in the ComOps department. A family member of mine has worked with him on a few occasions and that’s how I got in contact with him.

My Java program will search a list of sources for a given keyword and write the output to a database sorted by relevance. The sources will include most consumer news sites and will also include a Twitter API as well as the option to add more sources. Pre-installed sources will include

· <http://bbc.com/>

· <http://foxnews.com/>

· <http://www.huffingtonpost.com/>

· <http://www.nbcnews.com/>

· <http://www.nytimes.com/>

· <http://www.wsj.com/>

· <http://www.usatoday.com/>

· <http://news.google.com/>

· <http://www.rollcall.com>

· <http://www.latimes.com>

· <http://www.wired.com>

· <http://www.cnn.com>

· <http://www.npr.org>

· <http://abcnews.go.com/>

· <https://www.usnews.com/news>

· <https://www.yahoo.com/news/>

· <https://www.ap.org/en-us/>

All sources will be searched recursively but will never read duplicate links. Loops and recursion will be an essential part needed for searching. It will need to write to a database, likely Derby. GUI will be a very large part of this as I will not be the one using it. It must be intuitive, easy to follow, and include a help menu. The sorting will be done in two stages: one, a ‘dumb’ classifier (in Java) will read the data and sort it by the number of times the search keyword appears, and two, a ‘smart’ classifier (in Python). The smart classifier will be in Python because of the wealth of libraries available to use. First, the program will start with a list of sources. Then, those sources will be read into a class that will explore each link, read the article it points to, and search it for the given keyword. If they keyword is found, it will write it to a file AND a Vector set. The Vector data structure will be used to hold all links of any relevance to the keyword. Vector will be used because part of the program will involve multithreading, and Vector is thread-safe. Then, the Python classifier will take that file as input and explore each link in the file. A Latent Dirichlet Allocation model will be used to sort the data based suspected relevance to the keyword. This will output to a database, and this database will be able to be accessed by the user but in only read-only mode. The user will then be able use that data however they wish.

**Pseudocode:**

**Extractor**

**public boolean searchPageLinks(String URL, int depth, String strUserId, String strPassword)**

Initialize authentication string to username+password

Initialize error messages to a blank string

If link hasn’t already been searched and depth is less than max depth and no stop flags have been thrown

Mark URL as already searched

Connect to the url and create a document

If search term appears more than once in the document

Count the number and add it to the list of matches

For every link on page

searchPageLinks with depth incremented by one

**Project outline:**

Open Program

Get login information and save to login credentials

Open the main pane and load the sources into a list

Get and save the search keyword and options

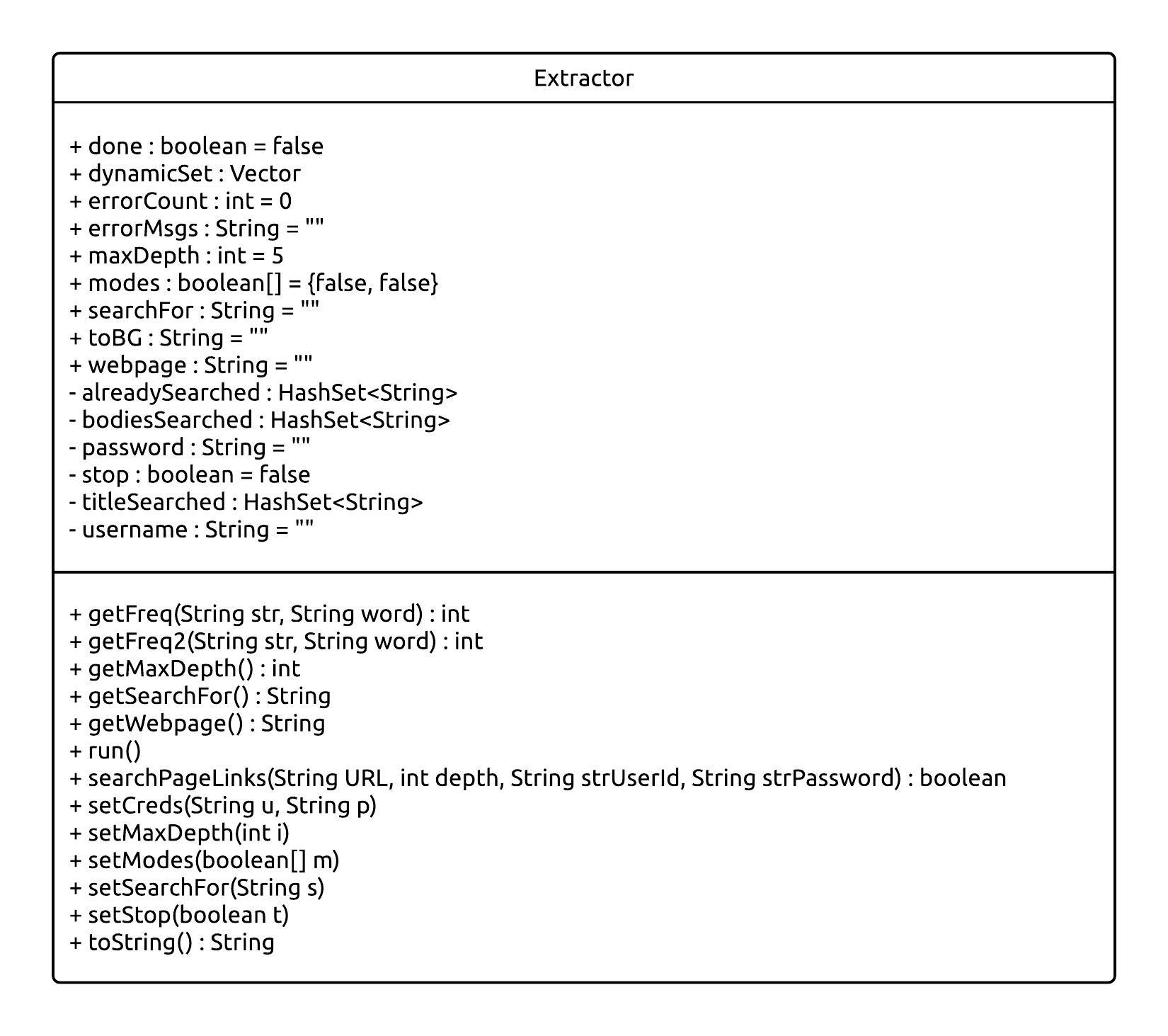
Sort the links

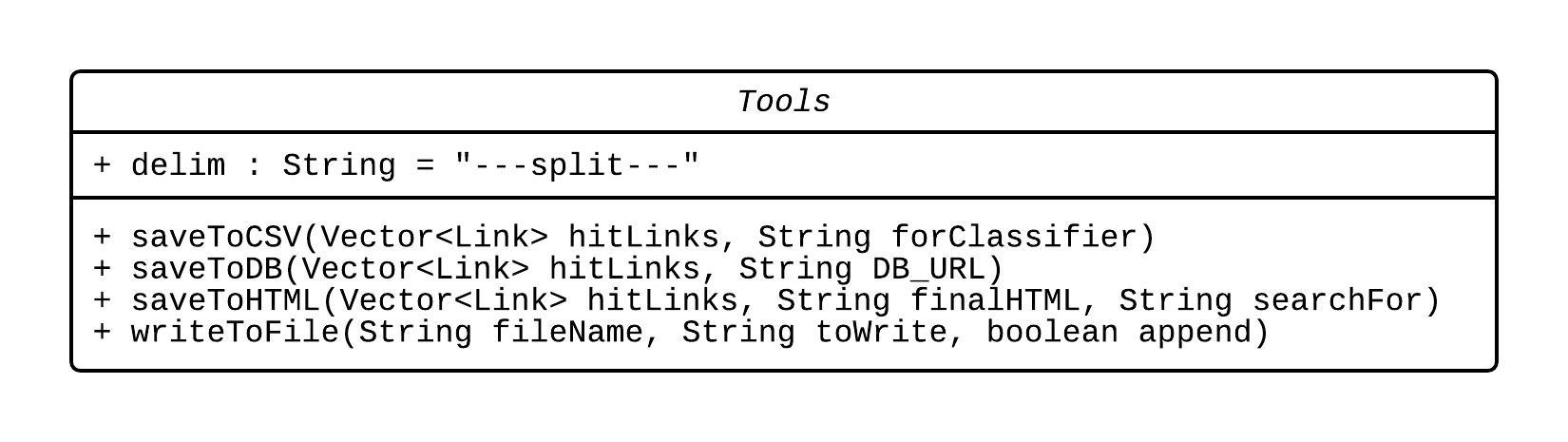
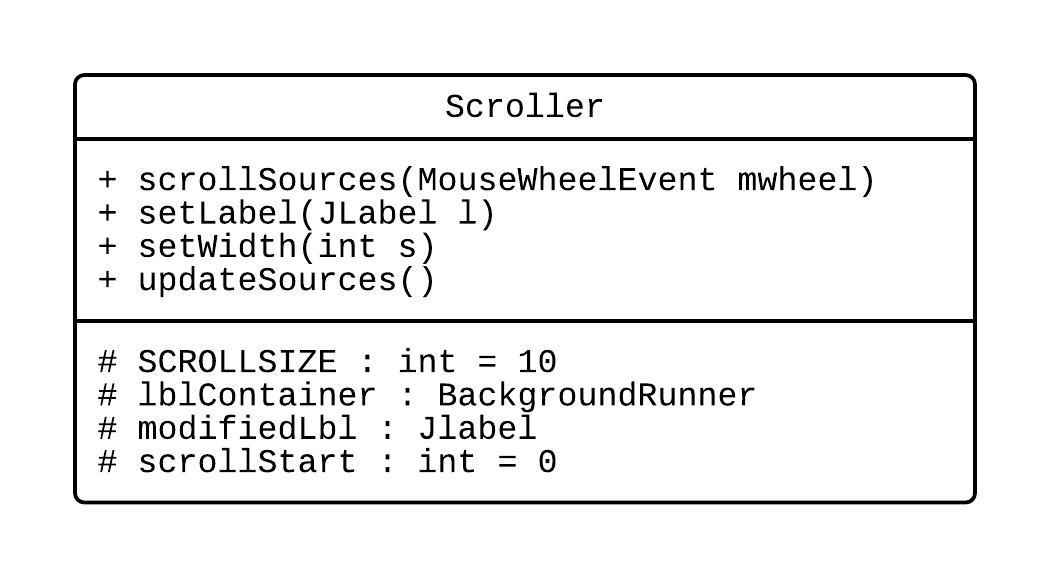
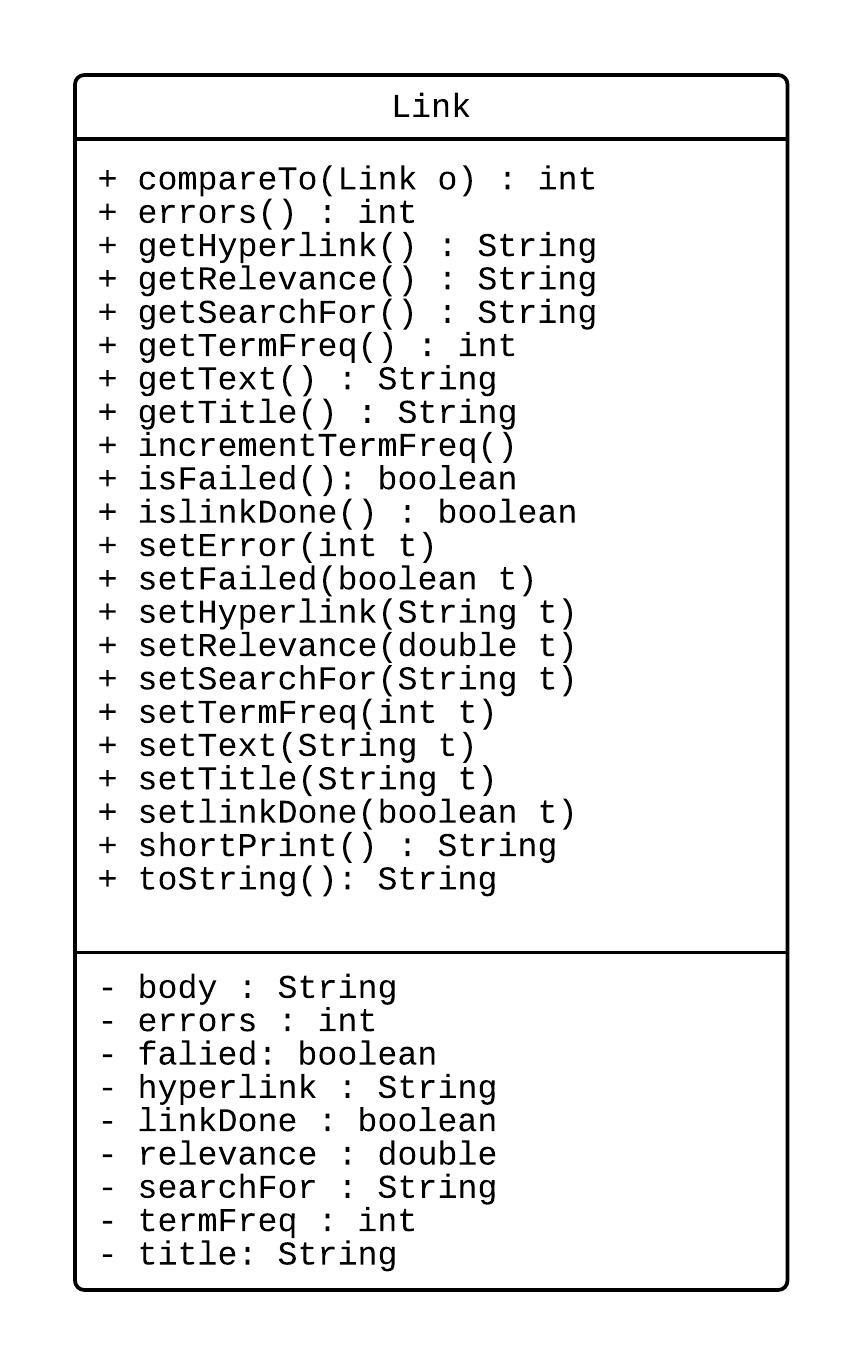
Save all links in order of relevance to database

**Test Cases:**

IA test cases:

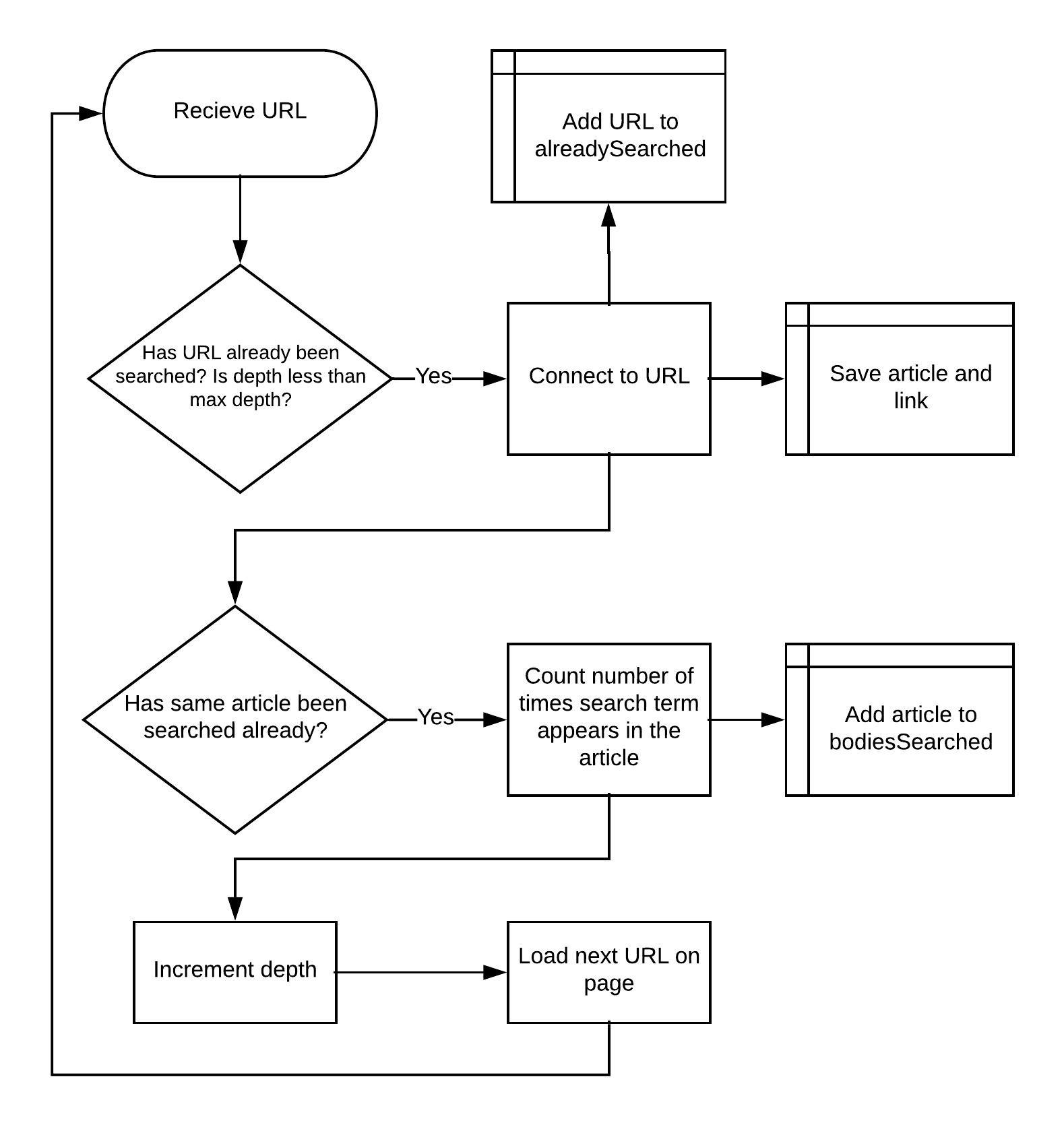
* **Reasonable**
  + Depth 2, Keyword ‘trump’
  + Depth 5, keyword ‘kavanaugh’
* **Boundary**
  + Depth 10, keyword ‘wave’
* **Extreme**
  + Depth 30, keyword ‘trump’
  + Depth 2, keyword ‘asdfsa’
* **Unreasonable**
  + Depth -1, keyword ‘asfd’

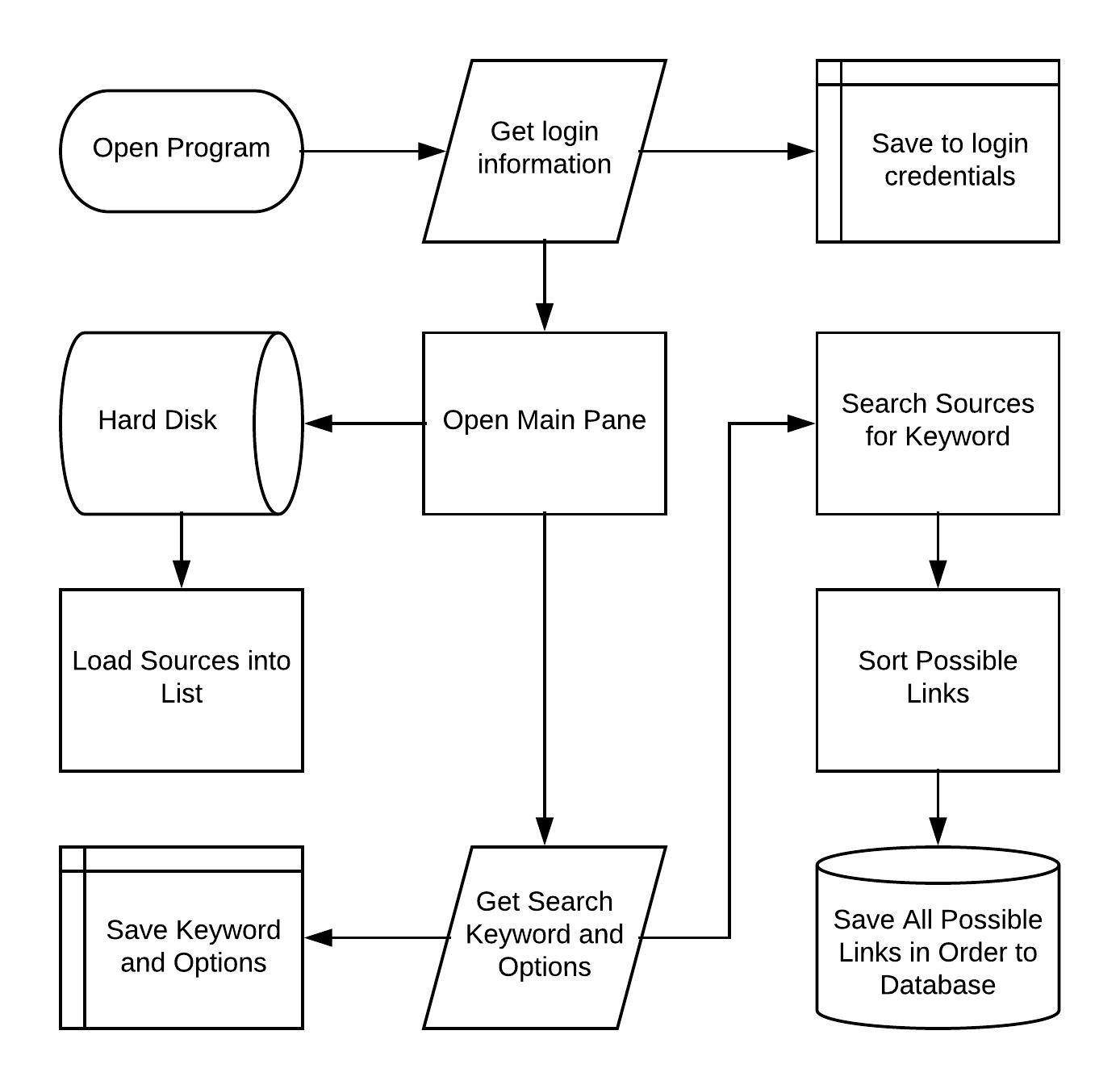
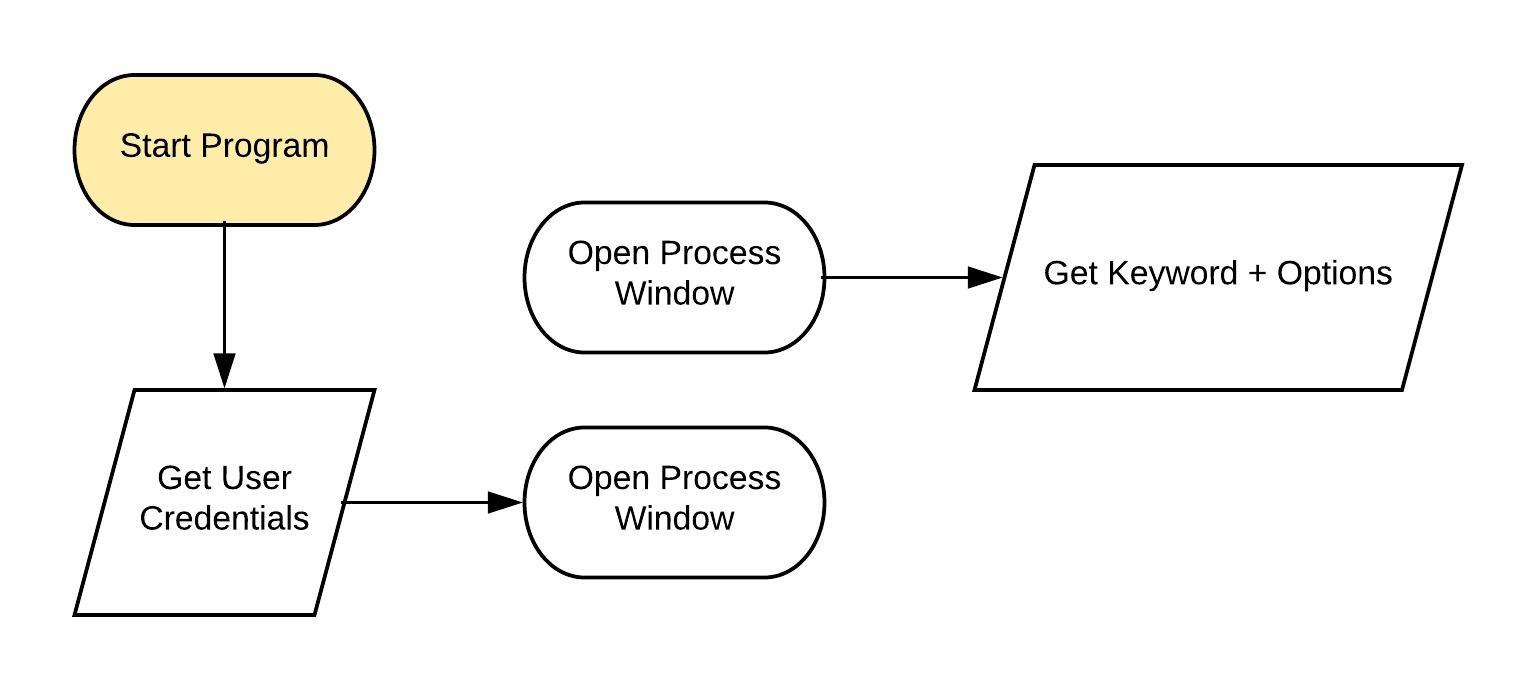
**UML Diagrams:**





**Flow Charts**

****



**Criterion B: Record of Tasks List**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Task  # | Planned action | Planned outcome | Time estimated | Target completion date | Criterion |
| **1** | **Meet with Joe** | **Get project idea.**  **Stimulate thought** | **short** | **Completed may 2018** |  |
| **2** | **Meet with Joe** | **Get plan**  **Get enough to work on**  **for a while** | **1 hour** | **Completed August 2018** |  |
| **3** | **Work on project** | **get work done**  **No goal, just as much**  **as possible** | **15 hrs** | **8/23-8/30** |  |
| **4** | **Make test cases** |  |  |  |  |
| **5** | **Test test cases** |  |  |  |  |
| **6** | **Flowchart** |  |  |  |  |
| **7** | **UML diagram** |  |  |  |  |
| **8** | **Write psuedocode** |  |  |  |  |
| **9** | **Write real code** |  |  |  |  |
| **10** | **Make movie** |  |  |  |  |
| **11** | **Get feedback from client** |  |  |  |  |
| **12** | **Write-up** |  |  |  |  |

**Criterion C: Interview and All Communication**

[Greetings]

Joe: Alright, so what is this school project.

Calvin: Well I’m taking IB computer science and we do a project with a client so I was wondering if you could maybe help me

Joe: Yes I probably could.

Calvin: Great! So do you have any problems you would need me to code a solution for?

Joe: Well we do a lot with news. We have commercial news sites that we use for getting things like the price of coal and stuff like that, but we don’t have anything to do that for consumer news sites. = **Problem**

Calvin: So what would that entail?

Joe: Well you would need to handle a lot of data and parse into links sorted by relevance. It would need to be stored in a database for the output. People on my team would use it so it would need to be easy to use. It would probably be a good idea to multithread because it’s a lot of data.

Calvin: Ok cool I’ll update you with my progress.

Joe: Sounds good.

[Calvin leaves]

Hi Calvin!

I hope you had a good summer! I spoke with my management earlier and we can only use data which is also available to the public, but that doesn’t mean we can’t still do something cool and useful to AEP. I had a couple of ideas, but I’m open to anything.

Com Ops uses a tremendous amount of information (for example news stories) from a large number of disparate sources. There is usually a lot of data presented that is of no use, but we still have to sift through it to find the few useful items. One idea I had was to write a program that could analyze various sites and produce a list of things we might be interested in. The most fun option might be to use the free (standard version) Twitter API (<https://developer.twitter.com/en/docs.html>), but you could always just pull in a couple of news sources instead. You could use a string compare or regex to find specific strings that would be of interest. Regular expressions are a little scary at first, but they are useful for a lot of different situations and worth learning. You mentioned that you were interested in machine learning, if you want to take that on you could use automatic classification tools (that will be a lot of work though). If you wanted to build a GUI, the user could input keywords to look for. You could write the results to an HTML file (maybe just links and a subject line) or even a database if you want to get ambitious. Just one idea, and if you like it we can flesh out the requirements.

Another option I thought of is to use the publicly available Data Miner 2 from the PJM website (<https://www.pjm.com/markets-and-operations/etools/data-miner-2.aspx>). PJM is a bulk electricity market that AEP sells our generation (energy created from our power plants) into and buys our load (the electricity our customers actually use). We can go over what the market is and isn’t if you want to go this route. Data Miner 2 has a web service API (<https://www.pjm.com/-/media/etools/data-miner-2/data-miner-2-api-guide.ashx?la=en>), though I’ve never used it so I’m not sure how complicated it is. Pretty much anything publicly available on Data Miner 2 would be useful. If you wanted you could store it into a database like derby (<https://db.apache.org/derby/>) or something.

Let me know what you think. I’m open to just about anything.

Thanks!

-Joe

Thanks for working to let me help you with some programs. I'm not really sure what the second option's goal would be, do you think you could elaborate a little on it? The first option looks like it would work well for me though, and I've done some regex stuff before in linux although I'll need to brush up a little on using that with Java. I also did some GUI work last year for a project so I could make a front end for it too. I'm interested in the ML automatic classification feature for that as well, and it would be really great to pipe the output to a database. What are the requirements for that project?

Also, when would you need me to come down to AEP next?

Thanks,

Calvin Kinateder

**Appendix A - Git Log**

85fe79f - (HEAD -> master, origin/master) removed unused file variable (10 days ago) <calvin kinateder>

7e5b41f - cleaned up (4 weeks ago) <Calvin>

4144fc6 - added op.gif + deleted some duplicate files (4 weeks ago) <Calvin>

b38b175 - sync (5 weeks ago) <Calvin>

2309db1 - fixed output not opening (5 weeks ago) <calvin kinateder>

41072a9 - removed 1 thing (5 weeks ago) <calvin kinateder>

734dbc9 - added more umls (5 weeks ago) <Calvin>

9ce9c2b - sync (5 weeks ago) <Calvin>

59fbce3 - fixed width issues (5 weeks ago) <calvin kinateder>

e3c7cd5 - output looks nicer also fixed the sorting issue by making a temp copy of the hitLinks (5 weeks ago) <calvin kinateder>

c18dea2 - changed to LDA multicore (5 weeks ago) <Calvin>

33a2253 - syncing classifier2 still crashes on CoherenceModel (5 weeks ago) <calvin kinateder>

e651149 - added mallet and trump3078 dataset (6 weeks ago) <Calvin>

6123ab1 - biggest dataset (6 weeks ago) <calvin kinateder>

91e44f2 - html works well (6 weeks ago) <Calvin>

22b658b - updated links.csv (6 weeks ago) <calvin kinateder>

4ecd6de - started saveToHTML() (6 weeks ago) <Calvin>

b6cbb72 - added getFreq2() (6 weeks ago) <Calvin>

ed6e59c - updated classifier.py (6 weeks ago) <calvin kinateder>

cd79d34 - classifier updated and moved some stuff around (6 weeks ago) <Calvin>

45ec391 - updated classifier.py (6 weeks ago) <calvin kinateder>

51a1805 - topic modeling ACTUALLY KIND OF WORKS!!! (7 weeks ago) <calvin kinateder>

341f3ff - no duplicate articles (7 weeks ago) <calvin kinateder>

20f172a - changed screenshots (7 weeks ago) <calvin kinateder>

0d42728 - changed a lot - changed datastructure and started on python (7 weeks ago) <calvin kinateder>

ec2f079 - sorting algorithm updated with multiple delimiters (7 weeks ago) <calvin kinateder>

cd31b0b - database works, added compareTo to Link, but sorting algorithm isn't good (7 weeks ago) <Calvin>

a98ea26 - added FinalLinks Derby database (7 weeks ago) <Calvin>

0c716be - tried to fix continuous errors on output (8 weeks ago) <Calvin>

0546791 - started migrating to ranking output by termFreq (8 weeks ago) <calvin kinateder>

e9d0b32 - update ia (9 weeks ago) <calvin kinateder>

51bc9bc - started framework for modes array (9 weeks ago) <Calvin>

530d7c6 - update ia (9 weeks ago) <calvin kinateder>

0f9e671 - hit count works! (9 weeks ago) <calvin kinateder>

5db03fb - added error color codes (9 weeks ago) <calvin kinateder>

962c7dd - added search settings and error color flashing (9 weeks ago) <calvin kinateder>

15562d5 - CHANGED SOURCES TO VECTOR (9 weeks ago) <Calvin>

47d0522 - Merge origin/master (9 weeks ago) <Calvin>

d7de0f5 - added link and stuff (9 weeks ago) <calvin kinateder>

a1e415c - delimit in comma csv (9 weeks ago) <calvin kinateder>

49944bb - Delete Mon Oct 8 13:49:14 EDT 2018.txt (9 weeks ago) <Calvin Kinateder>

6d92b7d - added purposes to each file, 10 (9 weeks ago) <Calvin>

f37ad39 - added log\_MAC.command (9 weeks ago) <Calvin>

b783566 - fixed scroller not updating and added catch to prohibit source modification while running (9 weeks ago) <Calvin>

0a75026 - updated working tree (10 weeks ago) <calvin kinateder>

39f540e - updated IA.docx (10 weeks ago) <calvin kinateder>

55dd6fd - fixed some formatting (10 weeks ago) <calvin kinateder>

30c34e2 - fixed some formatting (10 weeks ago) <calvin kinateder>

cfaec09 - about to migrate to full on LINK (10 weeks ago) <calvin kinateder>

2de8e05 - open out opens links.txt (10 weeks ago) <calvin kinateder>

bac4759 - ICONS WORK!!! (10 weeks ago) <calvin kinateder>

2637835 - mainly just updated working tree (10 weeks ago) <Calvin>

0d9b3d1 - added tooltips (2 months ago) <Calvin>

f90af2b - changed screenshot (2 months ago) <Calvin>

b7409a5 - made some things more consistent (2 months ago) <Calvin>

3b0a398 - renamed variables and took out the trash (2 months ago) <Calvin>

3959af7 - update (2 months ago) <calvin>

388ebbd - improved quit time by 99% (2 months ago) <Calvin>

60db300 - source editor doesnt work but ok to revert to (2 months ago) <Calvin>

b45ce35 - addasource sucks so were gonna fix that next (2 months ago) <calvin>

002011f - fixed scrolling stuff (2 months ago) <calvin>

7e16521 - more cosmetic (2 months ago) <Calvin>

b9d0c56 - added constant for output size and formatted scroller a little (2 months ago) <Calvin>

37f1337 - Added automatic refresh plus complete notification (3 months ago) <calvin>

dd443f3 - minor (3 months ago) <calvin>

2881453 - BIG ONE : THREAD STOP WORKS (3 months ago) <calvin>

d6c031f - modified some names (3 months ago) <Calvin>

b5d8ef1 - added uml project (3 months ago) <Calvin>

1448d59 - updated writeup (3 months ago) <Calvin>

b296504 - added reuters (3 months ago) <Calvin>

a5be823 - Added apache common codec (3 months ago) <Calvin>

607ff35 - change to vector (3 months ago) <calvin kinateder>

420dd88 - continuing linkset (3 months ago) <Calvin>

2031f8e - started linkset (3 months ago) <Calvin>

c808764 - Added Link Class (3 months ago) <Calvin>

a659979 - ADDED ia (3 months ago) <Calvin>

476a949 - replace datasets (3 months ago) <calvin kinateder>

ffde864 - added files (3 months ago) <Calvin>

8745104 - fixed source updating problem and added datasets (3 months ago) <Calvin>

471f20b - update sc (3 months ago) <Calvin>

84b7289 - added verbose option (3 months ago) <Calvin>

9e66540 - updated scroller (3 months ago) <calvin kinateder>

8608e6f - added long out (3 months ago) <calvin kinateder>

b44c9e6 - update (3 months ago) <Calvin>

1802279 - moved scroll to scroller (3 months ago) <Calvin>

2d68635 - disregard source changes, but scroll works (3 months ago) <Calvin>

435771e - finished message is good (3 months ago) <Calvin>

58bbc80 - fixed concurrency error when clicking extract when running (3 months ago) <Calvin>

37bea48 - MAJOR MAJOR MAJOR - STOP FINALLY WORKS OKAY - some sync probs tho (3 months ago) <calvin kinateder>

7dd70ea - update (3 months ago) <Calvin>

dfaec7b - progress output works well (3 months ago) <Calvin>

76a7beb - Updated formatting (3 months ago) <Calvin>

4c942a6 - Extends thread not implements runnable. now can access extractor vars so cool (3 months ago) <Calvin>

d467dbf - OK to revert - progress bar works great (3 months ago) <Calvin>

ea2f66e - dont revert just updating progress (3 months ago) <Calvin>

f48b8d5 - MAJOR. REVERT TO HERE. Progress works! (3 months ago) <Calvin>

3e489c3 - background runner works well (3 months ago) <Calvin>

51cf4cf - extractorcontainer is replaced by background runner (3 months ago) <Calvin>

49961df - just backing up work DO NOT REVERT (3 months ago) <Calvin>

7f73529 - update working tree dont revert to here (3 months ago) <calvin kinateder>

644b6e7 - wouldn't necessarily revert to here, just updating tree for desktop (3 months ago) <Calvin>

30c971f - modified props and little ui change (3 months ago) <Calvin>

3daad9b - added sources and authentication (3 months ago) <Calvin>

97c3158 - added refresh src button (3 months ago) <calvin kinateder>

d2dd75a - fixed readme.md (3 months ago) <calvin kinateder>

299b280 - Delete sgmain.PNG (3 months ago) <Calvin Kinateder>

30caa0f - Update README.md (3 months ago) <Calvin Kinateder>

9d361cb - fixed readme.md (3 months ago) <calvin kinateder>

98e4982 - fixed readme.md (3 months ago) <calvin kinateder>

bf4ef6f - fixed readme.md (3 months ago) <calvin kinateder>

13214a7 - revisited main layout (3 months ago) <calvin kinateder>

de6b00d - works about to change executor to larger scope ok to revert to (3 months ago) <calvin kinateder>

5c88b89 - about to change to executor.submit - OK TO REVERT (3 months ago) <calvin kinateder>

0a3ba6e - updated ok to revert (3 months ago) <calvin kinateder>

d5edead - source editor works ok to revert still needs work though (3 months ago) <calvin kinateder>

fc762cd - executerservice works just like thread rn. but added sources. multithreaded speed confirmed. (3 months ago) <calvin kinateder>

d212677 - Added some functions for joining threads, still stuck. beginning port to executor service. ok to revert to (3 months ago) <calvin kinateder>

d4fd45c - search works well now extract with threads still a little glitchy ok to revert to i guess (4 months ago) <calvin kinateder>

d6ffb98 - multithread works but still dont revert yet (4 months ago) <calvin kinateder>

4ff8515 - DO NOT REVERT TO THIS COMMIT -- BETA -- ADDED EXTRACTOR CONTAINER (4 months ago) <Calvin>

c6578f6 - single threaded works (4 months ago) <Calvin>

daa5669 - added twitter custom checkbox (4 months ago) <calvin kinateder>

e9d3cc9 - better code (4 months ago) <Calvin>

083a00c - attached remote javadoc to jtwitter: https://www.winterwell.com/software/jtwitter/javadoc/ (4 months ago) <Calvin>

f98f03b - added libs (4 months ago) <Calvin>

fd8ae4b - writes as it goes (4 months ago) <calvin kinateder>

e365048 - writes as it goes (4 months ago) <calvin kinateder>

a75edc0 - Changed main ui (4 months ago) <Calvin>

0b2ffd9 - Fixed broken project (4 months ago) <Calvin>

bc24ed3 - Fixed broken project (4 months ago) <Calvin>

ea2d69a - Delete project.xml (4 months ago) <Calvin Kinateder>

b96d615 - Delete project.properties (4 months ago) <Calvin Kinateder>

e53896a - Delete genfiles.properties (4 months ago) <Calvin Kinateder>

c5ee986 - Delete build-impl.xml (4 months ago) <Calvin Kinateder>

7ce2fd6 - Delete manifest.mf (4 months ago) <Calvin Kinateder>

4ac773d - Delete links.txt (4 months ago) <Calvin Kinateder>

b83606c - Delete build.xml (4 months ago) <Calvin Kinateder>

abbc215 - Delete acountss.txt (4 months ago) <Calvin Kinateder>

81ed1c3 - Delete Java 8 Articles (4 months ago) <Calvin Kinateder>

cd7cd4d - remove storygrab folder (4 months ago) <Calvin>

d2c516c - tst (4 months ago) <Calvin>

6509ccd - extractor works recursively (4 months ago) <Calvin>

32f04c5 - changed extractor (4 months ago) <Calvin>

8b9560b - added links (4 months ago) <Calvin>

32928d7 - added nonfunct xtractor (4 months ago) <calvin kinateder>

713fceb - added period (4 months ago) <calvin kinateder>

d0def41 - added help menu (4 months ago) <calvin kinateder>

bae6355 - fixed last commit (4 months ago) <Calvin>

1a828b3 - Update README.md (4 months ago) <Calvin Kinateder>

dd77a32 - Update README.md (4 months ago) <Calvin Kinateder>

ca3f5e8 - Update README.md (4 months ago) <Calvin Kinateder>

ed28358 - Merge origin/master (4 months ago) <Calvin>

bdc93de - BIG ONE: Removed access to create account panel-it was not necessary and it was insecure (4 months ago) <Calvin>

cb4d40b - Update README.md (4 months ago) <Calvin Kinateder>

0946599 - Update README.md (4 months ago) <Calvin Kinateder>

c58167b - Update README.md (4 months ago) <Calvin Kinateder>

627a131 - added screenshots (4 months ago) <calvin kinateder>

8eefafd - added two icons (4 months ago) <calvin kinateder>

d0aee4f - fixed corners by changing JFrame shape (4 months ago) <calvin kinateder>

0663cf6 - fixed mac issue with spawning pwdfield in create account (4 months ago) <Calvin>

2bb0652 - add logo (4 months ago) <calvin kinateder>

7e02be6 - change author from calvi to my actual name (4 months ago) <calvin kinateder>

3b0c9e6 - Cleaned up functions and improved create accounts screen (4 months ago) <calvi>

37f5a7e - Buttons and labels now work for release too (4 months ago) <calvi>

2a7cff4 - Big one: fixed buttons toggling and create account only works if it fields are filled all the way To do: pressed not clicked for buttons (4 months ago) <calvi>

a3ecd69 - added writing accounts to file and loading (4 months ago) <calvi>

694b1f3 - added create account (4 months ago) <calvi>

b159377 - update readme (4 months ago) <calvi>

81733bc - update readme (4 months ago) <calvi>

ce0b7fb - Added Accountmanager and user classes. (4 months ago) <Calvin>

bd1a92c - added checked and unchecked icons - may not be used (4 months ago) <Calvin>

94521e8 - login creates new screen (4 months ago) <Calvin>

3b0946e - Made it look nicer and added icons (4 months ago) <Calvin>

b55b0c3 - add lib (4 months ago) <calvi>

3030fbf - update readme (4 months ago) <calvi>

2e14007 - update readme (4 months ago) <calvi>

350a022 - Create README.md (4 months ago) <Calvin Kinateder>

abb3afc - Fixed opacity problem and added X (4 months ago) <calvi>

29493ca - First commit (4 months ago) <calvi>