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.

**Criterion B: Record of Tasks, Git History**

*55dd6fd 2018-10-06 | fixed some formatting (HEAD -> master, origin/master) [calvin kinateder]*

*30c34e2 2018-10-06 | fixed some formatting [calvin kinateder]*

*cfaec09 2018-10-06 | about to migrate to full on LINK [calvin kinateder]*

*2de8e05 2018-10-05 | open out opens links.txt [calvin kinateder]*

*bac4759 2018-10-05 | ICONS WORK!!! [calvin kinateder]*

*2637835 2018-10-05 | mainly just updated working tree [Calvin]*

*0d9b3d1 2018-10-04 | added tooltips [Calvin]*

*f90af2b 2018-10-04 | changed screenshot [Calvin]*

*b7409a5 2018-10-04 | made some things more consistent [Calvin]*

*3b0a398 2018-10-04 | renamed variables and took out the trash [Calvin]*

*3959af7 2018-10-03 | update [calvin]*

*388ebbd 2018-10-02 | improved quit time by 99% [Calvin]*

*60db300 2018-10-02 | source editor doesnt work but ok to revert to [Calvin]*

*b45ce35 2018-10-01 | addasource sucks so were gonna fix that next [calvin]*

*002011f 2018-10-01 | fixed scrolling stuff [calvin]*

*7e16521 2018-10-01 | more cosmetic [Calvin]*

*b9d0c56 2018-10-01 | added constant for output size and formatted scroller a little [Calvin]*

*37f1337 2018-09-29 | Added automatic refresh plus complete notification [calvin]*

*dd443f3 2018-09-28 | minor [calvin]*

*2881453 2018-09-28 | BIG ONE : THREAD STOP WORKS [calvin]*

*d6c031f 2018-09-28 | modified some names [Calvin]*

*b5d8ef1 2018-09-28 | added uml project [Calvin]*

*1448d59 2018-09-27 | updated writeup [Calvin]*

*b296504 2018-09-26 | added reuters [Calvin]*

*a5be823 2018-09-26 | Added apache common codec [Calvin]*

*607ff35 2018-09-25 | change to vector [calvin kinateder]*

*420dd88 2018-09-25 | continuing linkset [Calvin]*

*2031f8e 2018-09-25 | started linkset [Calvin]*

*c808764 2018-09-25 | Added Link Class [Calvin]*

*a659979 2018-09-24 | ADDED ia [Calvin]*

*476a949 2018-09-19 | replace datasets [calvin kinateder]*

*ffde864 2018-09-19 | added files [Calvin]*

*8745104 2018-09-18 | fixed source updating problem and added datasets [Calvin]*

*471f20b 2018-09-17 | update sc [Calvin]*

*84b7289 2018-09-17 | added verbose option [Calvin]*

*9e66540 2018-09-16 | updated scroller [calvin kinateder]*

*8608e6f 2018-09-14 | added long out [calvin kinateder]*

*b44c9e6 2018-09-14 | update [Calvin]*

*1802279 2018-09-14 | moved scroll to scroller [Calvin]*

*2d68635 2018-09-14 | disregard source changes, but scroll works [Calvin]*

*435771e 2018-09-14 | finished message is good [Calvin]*

*58bbc80 2018-09-13 | fixed concurrency error when clicking extract when running [Calvin]*

*37bea48 2018-09-13 | MAJOR MAJOR MAJOR - STOP FINALLY WORKS OKAY - some sync probs tho [calvin kinateder]*

*7dd70ea 2018-09-13 | update [Calvin]*

*dfaec7b 2018-09-13 | progress output works well [Calvin]*

*76a7beb 2018-09-13 | Updated formatting [Calvin]*

*4c942a6 2018-09-12 | Extends thread not implements runnable. now can access extractor vars so cool [Calvin]*

*d467dbf 2018-09-11 | OK to revert - progress bar works great [Calvin]*

*ea2f66e 2018-09-10 | dont revert just updating progress [Calvin]*

*f48b8d5 2018-09-10 | MAJOR. REVERT TO HERE.* *Progress works! [Calvin]*

*3e489c3 2018-09-10 | background runner works well [Calvin]*

*51cf4cf 2018-09-10 | extractorcontainer is replaced by background runner [Calvin]*

*49961df 2018-09-10 | just backing up work DO NOT REVERT [Calvin]*

*7f73529 2018-09-10 | update working tree dont revert to here [calvin kinateder]*

*644b6e7 2018-09-08 | wouldn't necessarily revert to here, just updating tree for desktop [Calvin]*

*30c971f 2018-09-07 | modified props and little ui change [Calvin]*

*3daad9b 2018-09-05 | added sources and authentication [Calvin]*

*97c3158 2018-09-03 | added refresh src button [calvin kinateder]*

*d2dd75a 2018-09-03 | fixed readme.md [calvin kinateder]*

*299b280 2018-09-03 | Delete sgmain.PNG [Calvin Kinateder]*

*30caa0f 2018-09-03 | Update README.md [Calvin Kinateder]*

*9d361cb 2018-09-03 | fixed readme.md [calvin kinateder]*

*98e4982 2018-09-03 | fixed readme.md [calvin kinateder]*

*bf4ef6f 2018-09-03 | fixed readme.md [calvin kinateder]*

*13214a7 2018-09-03 | revisited main layout [calvin kinateder]*

*de6b00d 2018-09-03 | works about to change executor to larger scope ok to revert to [calvin kinateder]*

*5c88b89 2018-09-02 | about to change to executor.submit - OK TO REVERT [calvin kinateder]*

*0a3ba6e 2018-09-01 | updated ok to revert [calvin kinateder]*

*d5edead 2018-09-01 | source editor works ok to revert still needs work though [calvin kinateder]*

*fc762cd 2018-09-01 | executerservice works just like thread rn. but added sources. multithreaded speed confirmed. [calvin kinateder]*

*d212677 2018-09-01 | Added some functions for joining threads, still stuck. beginning port to executor service. ok to revert to [calvin kinateder]*

*d4fd45c 2018-08-30 | search works well now extract with threads still a little glitchy ok to revert to i guess [calvin kinateder]*

*d6ffb98 2018-08-30 | multithread works but still dont revert yet [calvin kinateder]*

*4ff8515 2018-08-30 | DO NOT REVERT TO THIS COMMIT -- BETA -- ADDED EXTRACTOR CONTAINER [Calvin]*

*c6578f6 2018-08-30 | single threaded works [Calvin]*

*daa5669 2018-08-29 | added twitter custom checkbox [calvin kinateder]*

*e9d3cc9 2018-08-29 | better code [Calvin]*

*083a00c 2018-08-29 | attached remote javadoc to jtwitter: https://www.winterwell.com/software/jtwitter/javadoc/ [Calvin]*

*f98f03b 2018-08-29 | added libs [Calvin]*

*fd8ae4b 2018-08-28 | writes as it goes [calvin kinateder]*

*e365048 2018-08-28 | writes as it goes [calvin kinateder]*

*a75edc0 2018-08-28 | Changed main ui [Calvin]*

*0b2ffd9 2018-08-28 | Fixed broken project [Calvin]*

*bc24ed3 2018-08-28 | Fixed broken project [Calvin]*

*ea2d69a 2018-08-28 | Delete project.xml [Calvin Kinateder]*

*b96d615 2018-08-28 | Delete project.properties [Calvin Kinateder]*

*e53896a 2018-08-28 | Delete genfiles.properties [Calvin Kinateder]*

*c5ee986 2018-08-28 | Delete build-impl.xml [Calvin Kinateder]*

*7ce2fd6 2018-08-28 | Delete manifest.mf [Calvin Kinateder]*

*4ac773d 2018-08-28 | Delete links.txt [Calvin Kinateder]*

*b83606c 2018-08-28 | Delete build.xml [Calvin Kinateder]*

*abbc215 2018-08-28 | Delete acountss.txt [Calvin Kinateder]*

*81ed1c3 2018-08-28 | Delete Java 8 Articles [Calvin Kinateder]*

*cd7cd4d 2018-08-28 | remove storygrab folder [Calvin]*

*d2c516c 2018-08-28 | tst [Calvin]*

*6509ccd 2018-08-28 | extractor works recursively [Calvin]*

*32f04c5 2018-08-28 | changed extractor [Calvin]*

*8b9560b 2018-08-28 | added links [Calvin]*

*32928d7 2018-08-27 | added nonfunct xtractor [calvin kinateder]*

*713fceb 2018-08-27 | added period [calvin kinateder]*

*d0def41 2018-08-27 | added help menu [calvin kinateder]*

*bae6355 2018-08-27 | fixed last commit [Calvin]*

*1a828b3 2018-08-27 | Update README.md [Calvin Kinateder]*

*dd77a32 2018-08-27 | Update README.md [Calvin Kinateder]*

*ca3f5e8 2018-08-27 | Update README.md [Calvin Kinateder]*

*ed28358 2018-08-27 | Merge origin/master [Calvin]*

*bdc93de 2018-08-27 | BIG ONE: Removed access to create account panel-it was not necessary and it was insecure [Calvin]*

*cb4d40b 2018-08-27 | Update README.md [Calvin Kinateder]*

*0946599 2018-08-27 | Update README.md [Calvin Kinateder]*

*c58167b 2018-08-26 | Update README.md [Calvin Kinateder]*

*627a131 2018-08-26 | added screenshots [calvin kinateder]*

*8eefafd 2018-08-26 | added two icons [calvin kinateder]*

*d0aee4f 2018-08-25 | fixed corners by changing JFrame shape [calvin kinateder]*

*0663cf6 2018-08-25 | fixed mac issue with spawning pwdfield in create account [Calvin]*

*2bb0652 2018-08-25 | add logo [calvin kinateder]*

*7e02be6 2018-08-25 | change author from calvi to my actual name [calvin kinateder]*

*3b0c9e6 2018-08-25 | Cleaned up functions and improved create accounts screen [calvi]*

*37f5a7e 2018-08-25 | Buttons and labels now work for release too [calvi]*

*2a7cff4 2018-08-25 | 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 [calvi]*

*a3ecd69 2018-08-24 | added writing accounts to file and loading [calvi]*

*694b1f3 2018-08-24 | added create account [calvi]*

*b159377 2018-08-24 | update readme [calvi]*

*81733bc 2018-08-24 | update readme [calvi]*

*ce0b7fb 2018-08-24 | Added Accountmanager and user classes. [Calvin]*

*bd1a92c 2018-08-24 | added checked and unchecked icons - may not be used [Calvin]*

*94521e8 2018-08-24 | login creates new screen [Calvin]*

*3b0946e 2018-08-24 | Made it look nicer and added icons [Calvin]*

*b55b0c3 2018-08-24 | add lib [calvi]*

*3030fbf 2018-08-24 | update readme [calvi]*

*2e14007 2018-08-24 | update readme [calvi]*

*350a022 2018-08-23 | Create README.md [Calvin Kinateder]*

*abb3afc 2018-08-23 | Fixed opacity problem and added X [calvi]*

*29493ca 2018-08-23 | First commit [calvi]*

**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