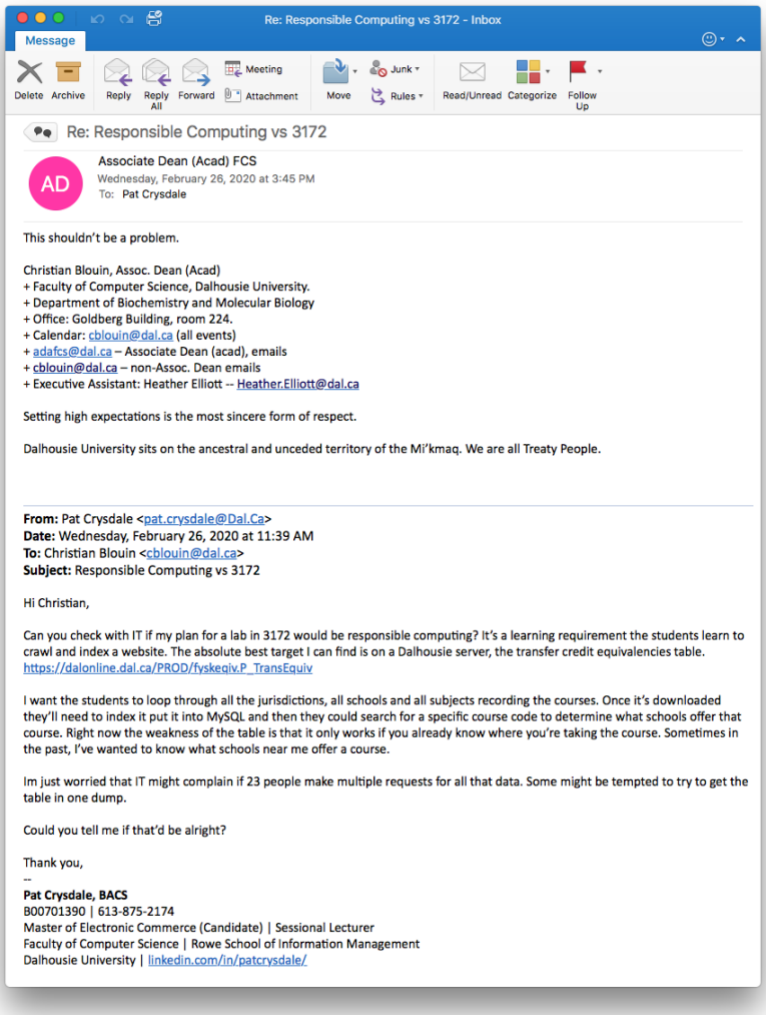
Lab 6 – Crawl, download, index, search and be awesome on all devices!

CSCI 3172 W20
Dalhousie University
Due: March 31th 11:59PM

This lab turned out to be huge and is worth 8% of your final grade.

In this lab, we will be crawling a Dalhousie webpage to get the contents of its database, and make it searchable. We have received special permission to crawl this specific resource from the Dean. It is always good practice to display such permission, so it is shown in the email below.



Re: Responsible Computing vs 3172 - Inbox

Message

Delete Archive Reply Reply All Forward Meeting Attachment Move Junk Rules Read/Unread Categorize Follow Up

Re: Responsible Computing vs 3172

AD

Associate Dean (Acad) FCS
Wednesday, February 26, 2020 at 3:45 PM
To: Pat Crysdale

This shouldn't be a problem.

Christian Blouin, Assoc. Dean (Acad)
+ Faculty of Computer Science, Dalhousie University.
+ Department of Biochemistry and Molecular Biology
+ Office: Goldberg Building, room 224.
+ Calendar: cblouin@dal.ca (all events)
+ adafcs@dal.ca – Associate Dean (acad), emails
+ cblouin@dal.ca – non-Assoc. Dean emails
+ Executive Assistant: Heather Elliott -- Heather.Elliott@dal.ca

Setting high expectations is the most sincere form of respect.

Dalhousie University sits on the ancestral and unceded territory of the Mi'kmaq. We are all Treaty People.

From: Pat Crysdale <pat.crysdale@Dal.Ca>
Date: Wednesday, February 26, 2020 at 11:39 AM
To: Christian Blouin <cblouin@dal.ca>
Subject: Responsible Computing vs 3172

Hi Christian,

Can you check with IT if my plan for a lab in 3172 would be responsible computing? It's a learning requirement the students learn to crawl and index a website. The absolute best target I can find is on a Dalhousie server, the transfer credit equivalencies table.
https://dalonline.dal.ca/PROD/fyskeqiv.P_TransEquiv

I want the students to loop through all the jurisdictions, all schools and all subjects recording the courses. Once it's downloaded they'll need to index it put it into MySQL and then they could search for a specific course code to determine what schools offer that course. Right now the weakness of the table is that it only works if you already know where you're taking the course. Sometimes in the past, I've wanted to know what schools near me offer a course.

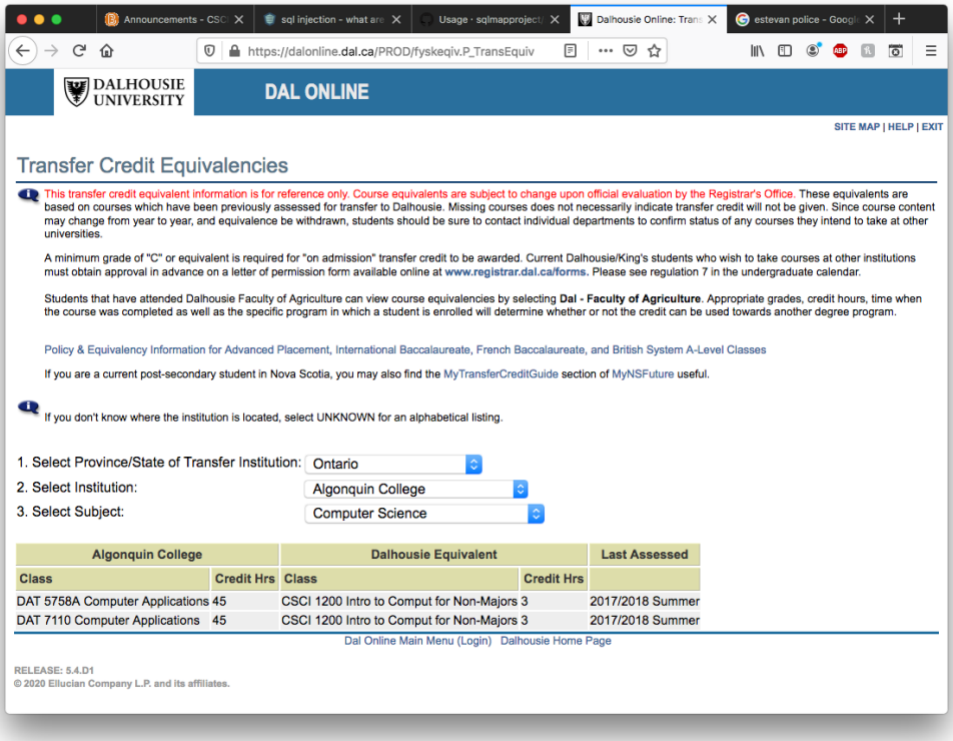
Im just worried that IT might complain if 23 people make multiple requests for all that data. Some might be tempted to try to get the table in one dump.

Could you tell me if that'd be alright?

Thank you,
--
Pat Crysdale, BACS
B00701390 | 613-875-2174
Master of Electronic Commerce (Candidate) | Sessional Lecturer
Faculty of Computer Science | Rowe School of Information Management
Dalhousie University | linkedin.com/in/patcrysdale/

The only resource that we are permitted to crawl is:
https://dalonline.dal.ca/PROD/fyskeqiv.P_TransEquiv

This resource is a database maintained by the registrar and contains previously approved transfer credits from other institutions. Shown below are Computer Science credits that transfer from Algonquin College.



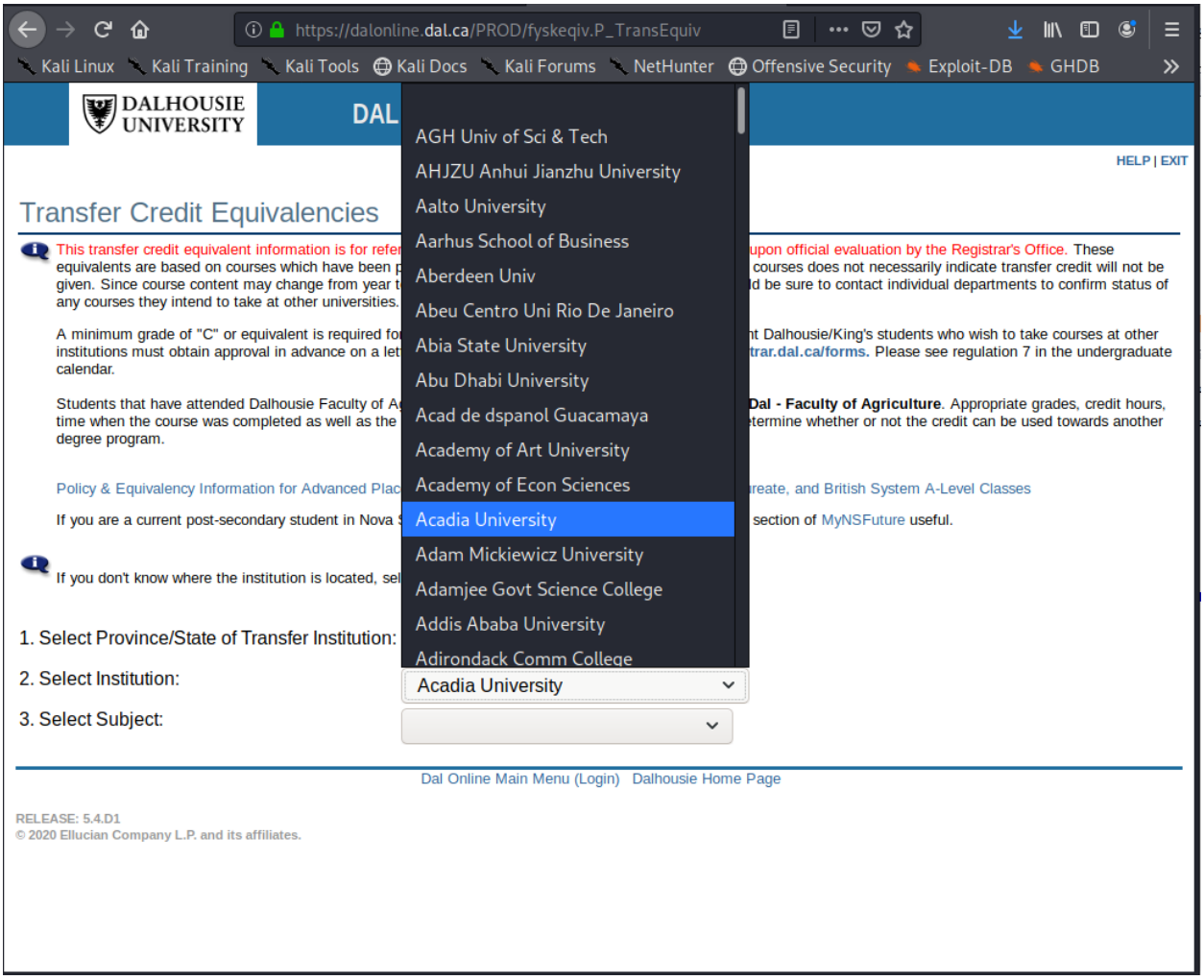
The only way to search this database is to know the name of the school and the department. You cannot search by course code. This isn't very user friendly Your job is to build an app with this functionality.

Steps to complete this lab:

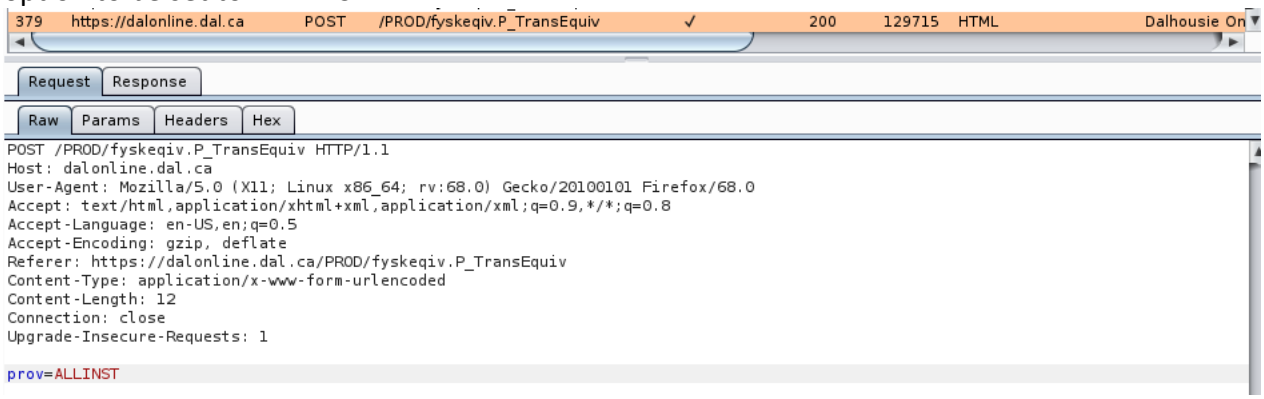
1. Crawl all subjects, at all institutions in all regions and download every table
2. Index the data and upload to your MySQL server
3. Create a webpage that has a search bar that can query your results by all headers. The website should be user friendly and responsive to desktop, tablet and mobile devices.

Do not worry. The basics of how to do all of the above are contained in the following pages. You will be provided with a significant shell to complete this lab.

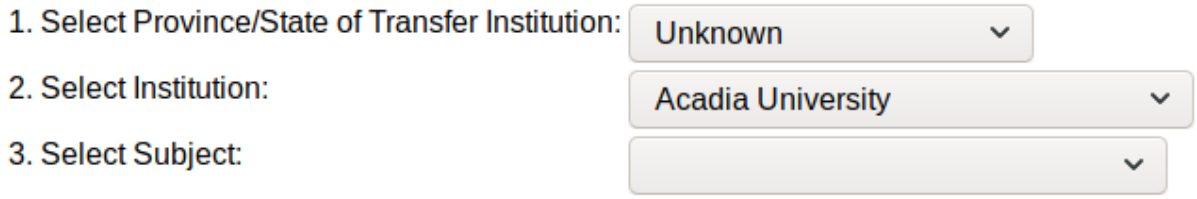
We will start by crawling the database to download every subject table, from every school. Choose “Unknown” in the first drop-down menu. This will put every school in the database in the institution field.



Passing through Burp we can see that there is a post parameter called “prov” which has an option to be set to “ALLINST”.



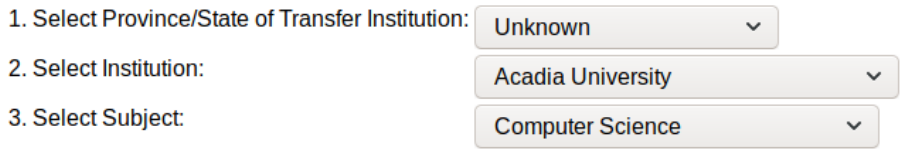
Choosing Acadia university (or any school) makes a subject field appear




```
POST /PROD/fyskeqiv.P_TransEquiv HTTP/1.1
Host: dalonline.dal.ca
User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:68.0) Gecko/20100101 Firefox/68.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
Referer: https://dalonline.dal.ca/PROD/fyskeqiv.P_TransEquiv
Content-Type: application/x-www-form-urlencoded
Content-Length: 24
Connection: close
Upgrade-Insecure-Requests: 1

prov=ALLINST&inst=NSU002
```

We see now that there's another post parameter called inst that specifies the school. Finally, let's pick a subject… burp says the last parameter is called subj

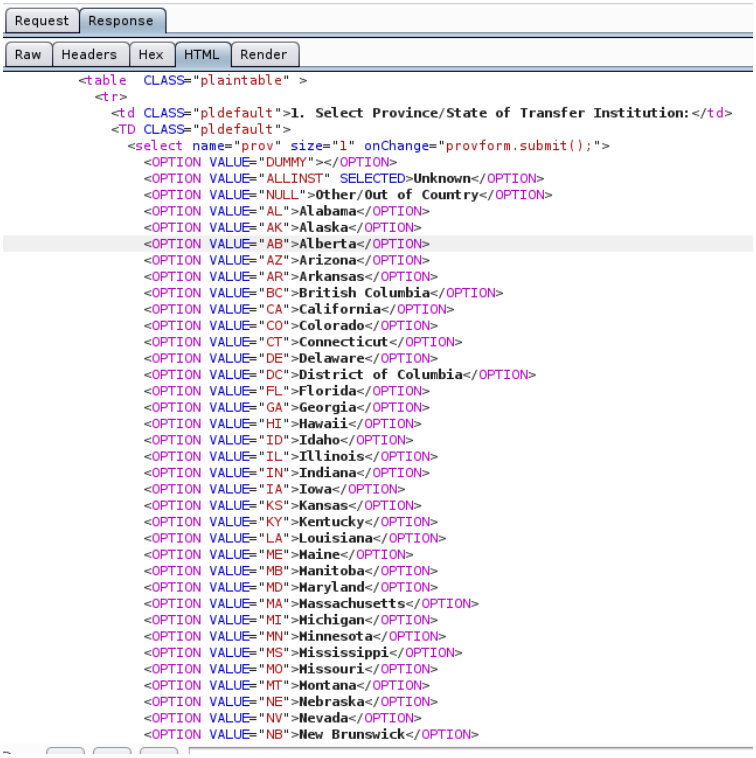


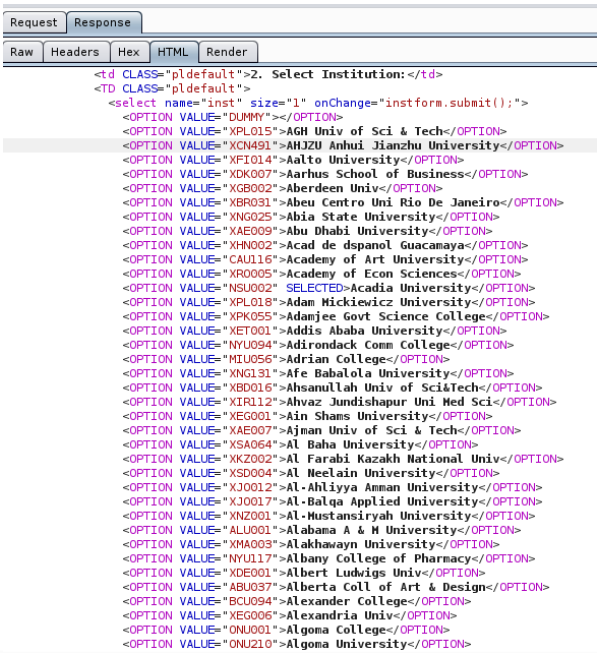
Acadia University		Dalhousie Equivalent		Last Assessed
Class	Credit Hrs	Class	Credit Hrs	
COMP 1813 Computer Concepts and App	3	CSCI 1200 Intro to Comput for Non-Majors	3	2010/2011 Fall
COMP 1863 Computer Prog & General Appl	3	CSCI 1105 Intro to Computer Programming	3	2009/2010 Winter
COMP 1893 Multimedia Based Intro: Prog	3	CSCI 1200 Intro to Comput for Non-Majors	3	2009/2010 Winter
COMP 2203 Computer Architecture/Organiza	3	CSCI 2999 2000 level elective	3	2013/2014 Winter
COMP 2213 Comp. Architect/Organization 2	3	CSCI 2991 2000 Level Elective	3	2019/2020 Fall
COMP 2853 Handling Data Concepts/Apps	3	CSCI 2999 2000 Level Elective	3	2018/2019 Summer
COMP 2863 How Web Sites Work		CSCI 1170 Intro to Web Design & Devel.	3	2019/2020 Fall
COMP 3403 Analysis of Algorithms	3	CSCI 3110 Dsgn. & Anal. of Algorithms I.	3	2017/2018 Summer

```
POST /PROD/fyskeqiv.P_TransEquiv HTTP/1.1
Host: dalonline.dal.ca
User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:68.0) Gecko/20100101 Firefox/68.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
Referer: https://dalonline.dal.ca/PROD/fyskeqiv.P_TransEquiv
Content-Type: application/x-www-form-urlencoded
Content-Length: 34
Connection: close
Upgrade-Insecure-Requests: 1

prov=ALLINST&inst=NSU002&subj=CSCI
```

We've now done enough information gathering to know that the table data is queried with: "prov, inst and subj". We now need values to give our crawler. These are stored in the dropdown menu, so we simply have to look at the source code. In Burp, look at under "Response" and then "HTML" for the HTTP packet you've intercepted.



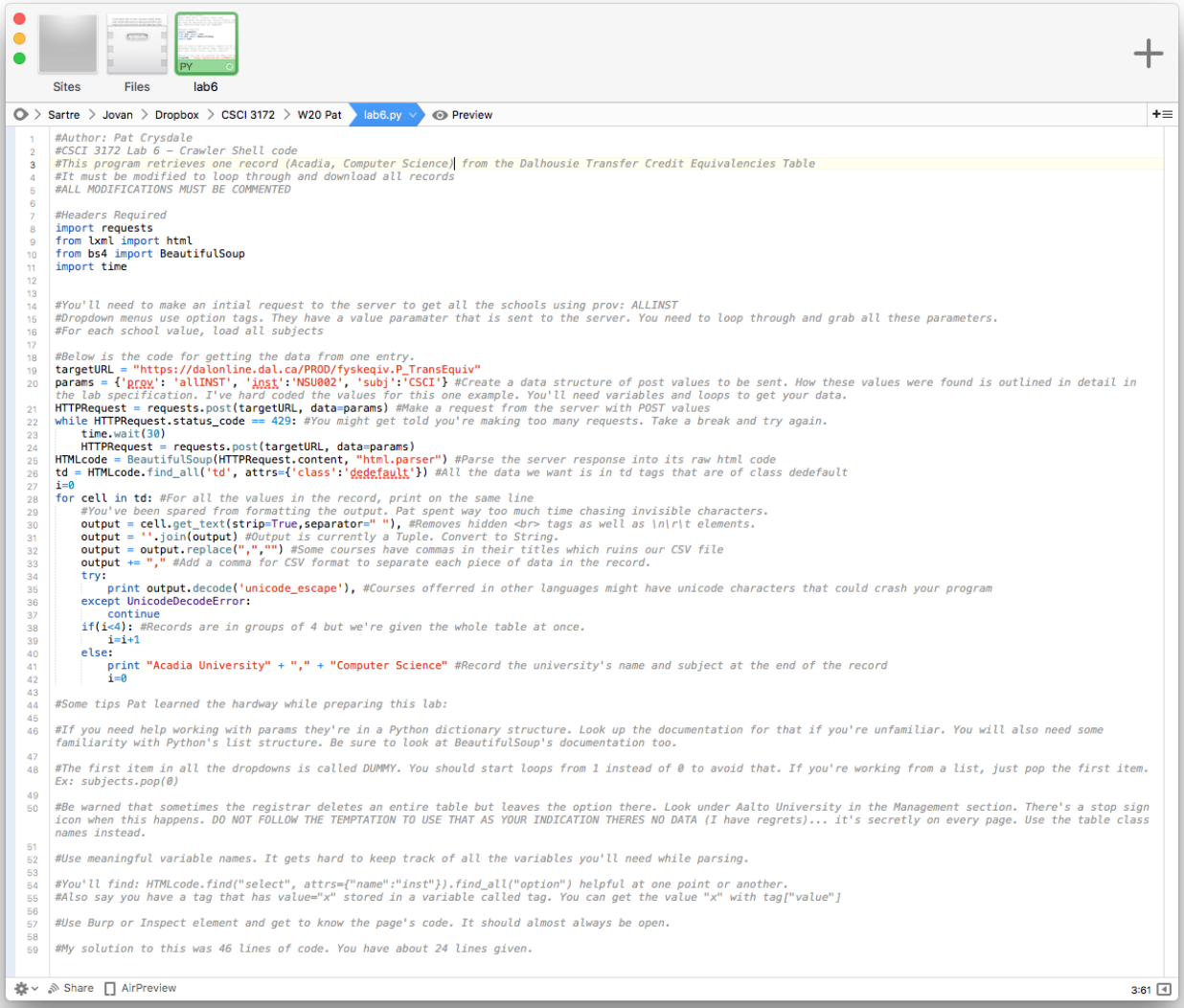


You will notice that all that data is there but is too much for a human. This is most definitely a job for a crawler.

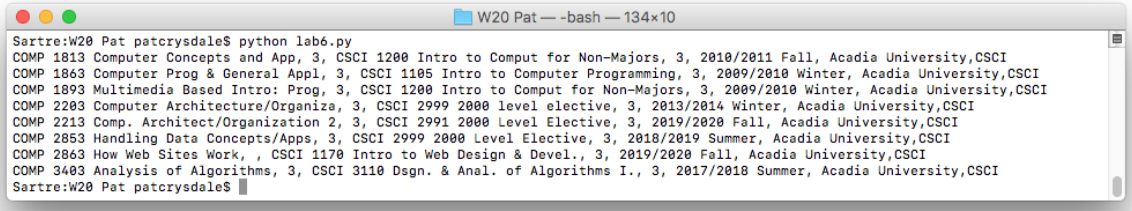
In class, I showed how to use Python's BeautifulSoup and Scapy libraries to crawl and gather information from sites. You will need one of those libraries to complete this lab.

I've provided a starting point on Brightspace called lab6.py. My program is thoroughly commented. This should be enough of an introduction that with some tinkering you should be able to complete it.

Here is a screen capture of that code.



Its output looks like this…

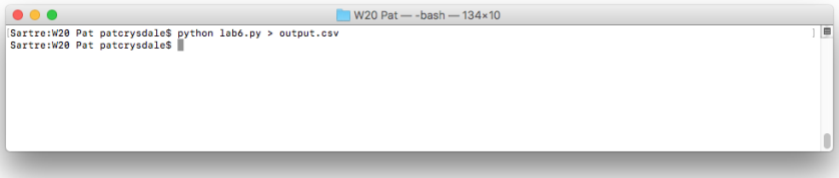


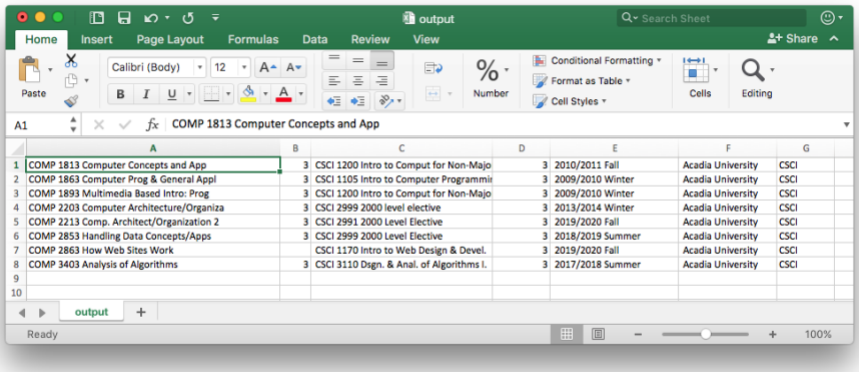
You have just extracted all of the data from this table. See?

1. Select Province/State of Transfer Institution: Unknown
2. Select Institution: Acadia University
3. Select Subject: Computer Science

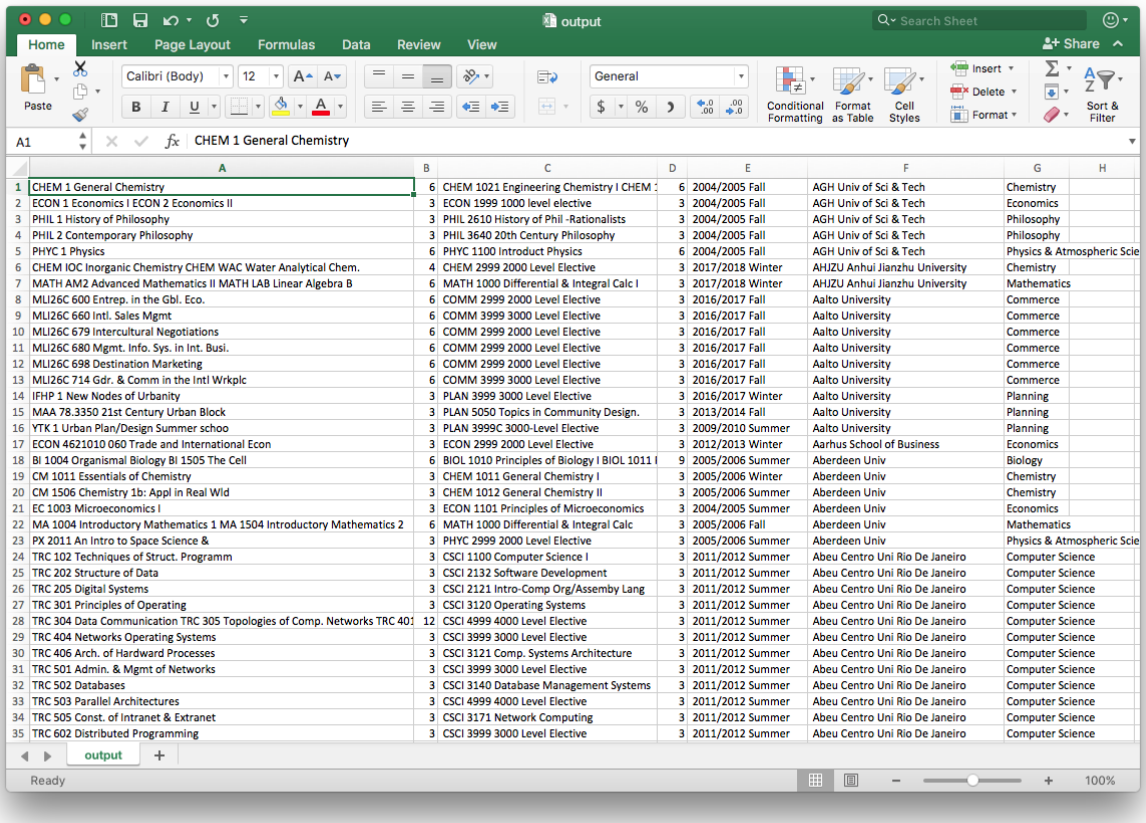
Acadia University		Dalhousie Equivalent		Last Assessed
Class	Credit Hrs	Class	Credit Hrs	
COMP 1813 Computer Concepts and App	3	CSCI 1200 Intro to Comput for Non-Majors	3	2010/2011 Fall
COMP 1863 Computer Prog & General Appl	3	CSCI 1105 Intro to Computer Programming	3	2009/2010 Winter
COMP 1893 Multimedia Based Intro: Prog	3	CSCI 1200 Intro to Comput for Non-Majors	3	2009/2010 Winter
COMP 2203 Computer Architecture/Organiza	3	CSCI 2999 2000 level elective	3	2013/2014 Winter
COMP 2213 Comp. Architect/Organization 2	3	CSCI 2991 2000 Level Elective	3	2019/2020 Fall
COMP 2853 Handling Data Concepts/Apps	3	CSCI 2999 2000 Level Elective	3	2018/2019 Summer
COMP 2863 How Web Sites Work		CSCI 1170 Intro to Web Design & Devel.	3	2019/2020 Fall
COMP 3403 Analysis of Algorithms	3	CSCI 3110 Dsgn. & Anal. of Algorithms I.	3	2017/2018 Summer

Write the output of your code to a .csv file. In your terminal write “python [filename].py > [someoutputfile].csv”.





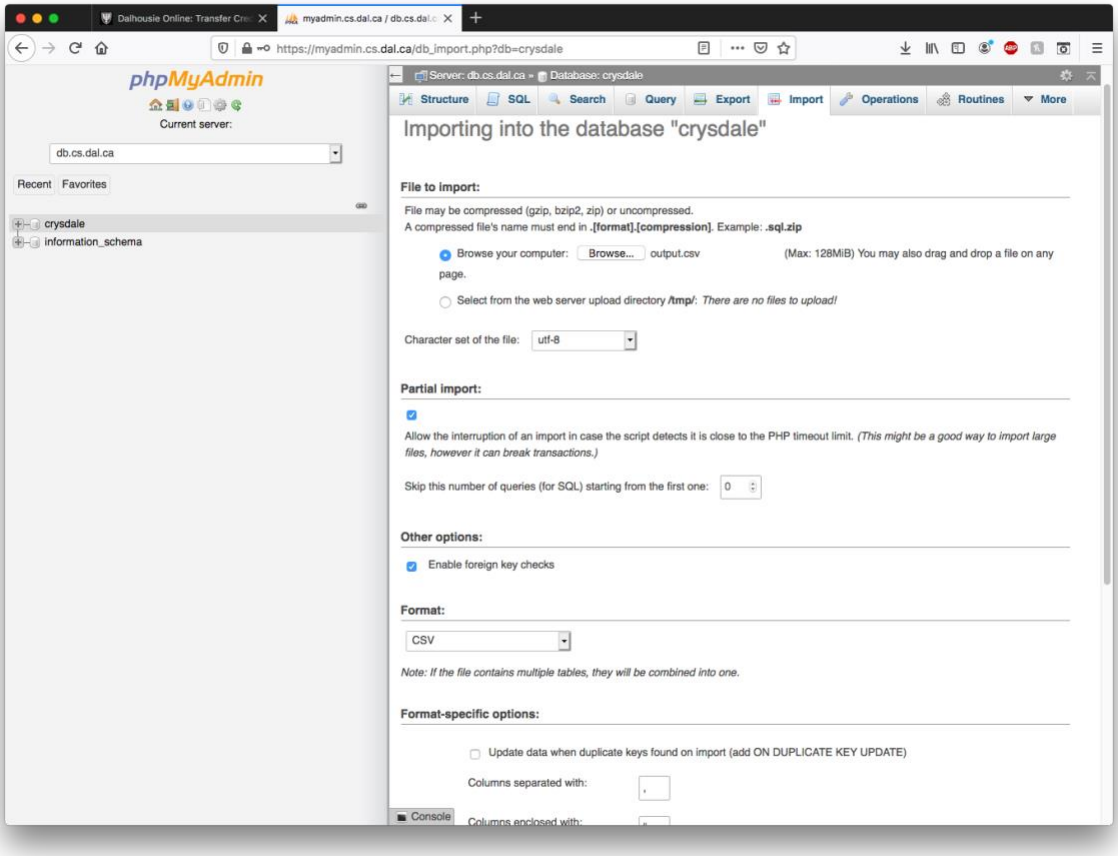
Here is a sample of what your dataset should look like once you've crawled the entire table.



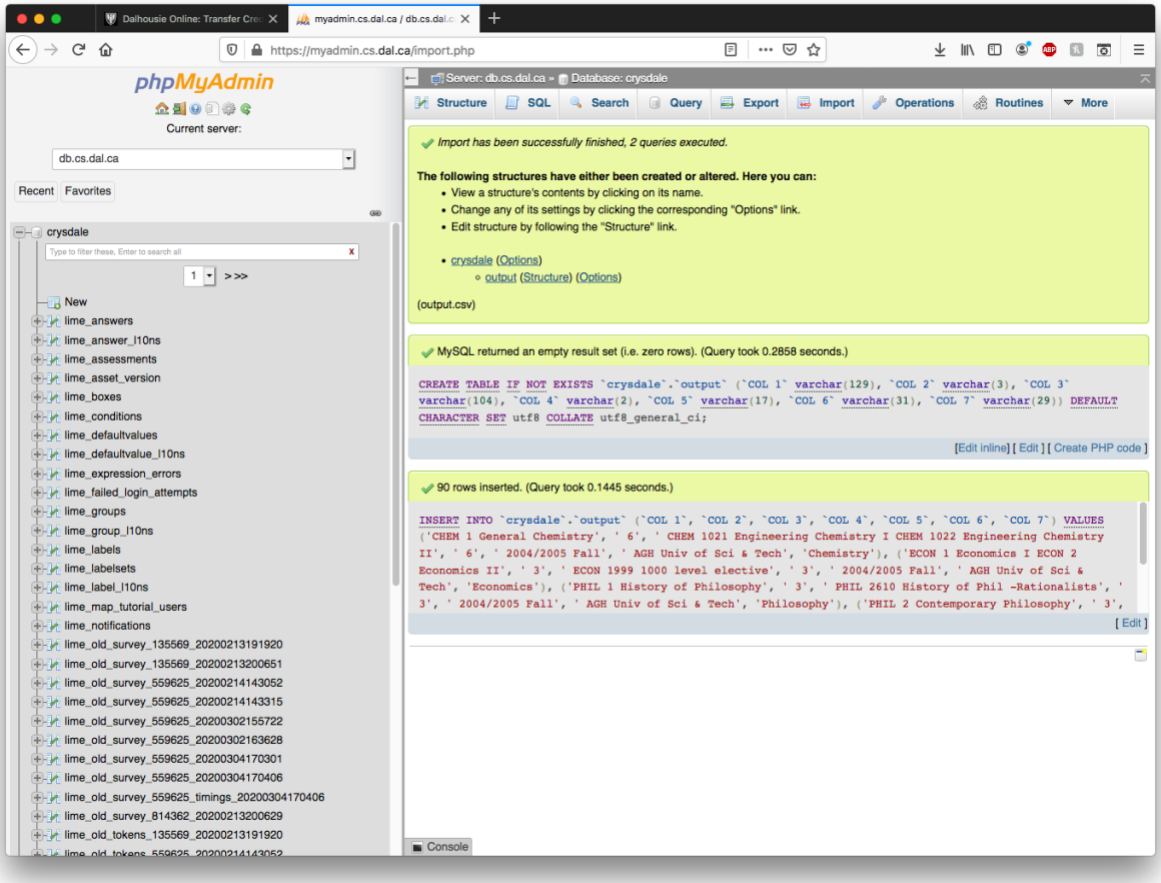
	A	B	C	D	E	F	G	H
1	CHEM 1 General Chemistry	6	CHEM 1021 Engineering Chemistry I CHEM 1	6	2004/2005 Fall	AGH Univ of Sci & Tech	Chemistry	
2	ECON 1 Economics I ECON 2 Economics II	3	ECON 1999 1000 level elective	3	2004/2005 Fall	AGH Univ of Sci & Tech	Economics	
3	PHIL 1 History of Philosophy	3	PHIL 2610 History of Phil -Rationalists	3	2004/2005 Fall	AGH Univ of Sci & Tech	Philosophy	
4	PHIL 2 Contemporary Philosophy	3	PHIL 3640 20th Century Philosophy	3	2004/2005 Fall	AGH Univ of Sci & Tech	Philosophy	
5	PHYC 1 Physics	6	PHYC 1100 Introduct Physics	6	2004/2005 Fall	AGH Univ of Sci & Tech	Physics & Atmospheric Scie	
6	CHEM IOC Inorganic Chemistry CHEM WAC Water Analytical Chem.	4	CHEM 2999 2000 Level Elective	3	2017/2018 Winter	AHJZU Anhui Jianzhu University	Chemistry	
7	MATH AM2 Advanced Mathematics II MATH LAB Linear Algebra B	6	MATH 1000 Differential & Integral Calc I	3	2017/2018 Winter	AHJZU Anhui Jianzhu University	Mathematics	
8	MLI26C 600 Entrep. in the Gbl. Eco.	6	COMM 2999 2000 Level Elective	3	2016/2017 Fall	Aalto University	Commerce	
9	MLI26C 660 Intl. Sales Mgmt	6	COMM 3999 3000 Level Elective	3	2016/2017 Fall	Aalto University	Commerce	
10	MLI26C 679 Intercultural Negotiations	6	COMM 2999 2000 Level Elective	3	2016/2017 Fall	Aalto University	Commerce	
11	MLI26C 680 Mgmt. Info. Sys. in Int. Busi.	6	COMM 2999 2000 Level Elective	3	2016/2017 Fall	Aalto University	Commerce	
12	MLI26C 698 Destination Marketing	6	COMM 2999 2000 Level Elective	3	2016/2017 Fall	Aalto University	Commerce	
13	MLI26C 714 Gdr. & Comm in the Intl Wrkplc	6	COMM 3999 3000 Level Elective	3	2016/2017 Fall	Aalto University	Commerce	
14	IFHP 1 New Nodes of Urbanity	3	PLAN 3999 3000 Level Elective	3	2016/2017 Winter	Aalto University	Planning	
15	MAA 78.3350 21st Century Urban Block	3	PLAN 5050 Topics in Community Design.	3	2013/2014 Fall	Aalto University	Planning	
16	YTK 1 Urban Plan/Design Summer schoo	3	PLAN 3999C 3000-Level Elective	3	2009/2010 Summer	Aalto University	Planning	
17	ECON 4621010 060 Trade and International Econ	3	ECON 2999 2000 Level Elective	3	2012/2013 Winter	Aarhus School of Business	Economics	
18	BI 1004 Organismal Biology BI 1505 The Cell	6	BIOL 1010 Principles of Biology I BIOL 1011 I	9	2005/2006 Summer	Aberdeen Univ	Biology	
19	CM 1011 Essentials of Chemistry	3	CHEM 1011 General Chemistry I	3	2005/2006 Winter	Aberdeen Univ	Chemistry	
20	CM 1506 Chemistry 1b: Appl in Real Wld	3	CHEM 1012 General Chemistry II	3	2005/2006 Summer	Aberdeen Univ	Chemistry	
21	EC 1003 Microeconomics I	3	ECON 1101 Principles of Microeconomics	3	2004/2005 Summer	Aberdeen Univ	Economics	
22	MA 1004 Introductory Mathematics 1 MA 1504 Introductory Mathematics 2	6	MATH 1000 Differential & Integral Calc	3	2005/2006 Fall	Aberdeen Univ	Mathematics	
23	PX 2011 An Intro to Space Science &	3	PHYC 2999 2000 Level Elective	3	2005/2006 Summer	Aberdeen Univ	Physics & Atmospheric Scie	
24	TRC 102 Techniques of Struct. Programm	3	CSCI 1100 Computer Science I	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
25	TRC 202 Structure of Data	3	CSCI 2132 Software Development	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
26	TRC 205 Digital Systems	3	CSCI 2121 Intro-Comp Org/Assemby Lang	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
27	TRC 301 Principles of Operating	3	CSCI 3120 Operating Systems	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
28	TRC 304 Data Communication TRC 305 Topologies of Comp. Networks TRC 401	12	CSCI 4999 4000 Level Elective	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
29	TRC 404 Networks Operating Systems	3	CSCI 3999 3000 Level Elective	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
30	TRC 406 Arch. of Hardward Processes	3	CSCI 3121 Comp. Systems Architecture	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
31	TRC 501 Admin. & Mgmt of Networks	3	CSCI 3999 3000 Level Elective	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
32	TRC 502 Databases	3	CSCI 3140 Database Management Systems	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
33	TRC 503 Parallel Architectures	3	CSCI 4999 4000 Level Elective	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
34	TRC 505 Const. of Intranet & Extranet	3	CSCI 3171 Network Computing	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	
35	TRC 602 Distributed Programming	3	CSCI 3999 3000 Level Elective	3	2011/2012 Summer	Abeu Centro Uni Rio De Janeiro	Computer Science	

Part 2:

You now have your data. If it's a CSV file it can be uploaded to your MySQL database.



You’re now ready to index your database. Do some research and find choose an indexing method that works for your needs.



PS You’re going to have more than 90 rows.

Part 3:

Grab your favorite front end framework and make a basic search page. Its wireframe should be similar to Google's. Your search page must be responsive and adapt to desktop, tablet and mobile interfaces.

Connect the search bar to your MySQL database.

Your search bar should be able to search and retrieve based on any combination of the following criteria:

- Host university
- Course name/code at host
- Course name/code at Dalhousie – (e.g. All equivalents to CSCI 3172)
- Credit hours
- Approval date

Output the results of your search.

Section	Not Submitted (0%)	Approaching Expectations (<80%)	Meeting Expectations (=80%)	Exceeds Expectations (>80%)
Crawler			All pages scraped CSV output is cleanly formatted Modifications to code have reasonable comments.	
MySQL database / Retrieved dataset			Data is indexed. Why you chose your indexing strategy is justified Entire dataset present.	
Search			Is responsive to mobile, tablet and desktop interfaces Searches entire dataset effectively. (TAs will be given random records to test). Search results are displayed in a usable format Search options are implemented (Can search for combination of terms)	