Lecuture Notes: Friday 30th September 2022

The professor started the class with reviewing material for the quiz. He mentioned that the material for the quiz will be the videos posted on canvas. Also requested the students to update any useful links found for studying sql on github.

The class further talked about Jobs database, viewing previous projects and introducing the idea of web scrapping to get data for projects. The professor while introducing us to project emphasized on the idea of how an ideal abstract for the project should look like.

Abstract should be direct to the point for example: “This is a notebook to scrape data for Finance jobs from scrapping indeed Finance. Professor demonstrated through a video on how to get multiple pages to scrape by changing url. He talked about learning python for web scrapping using loops(While loop) and how to stop or check if there are any more relevant pages.

**Web scrapping** using python useful links: <https://www.youtube.com/watch?v=XVv6mJpFOb0>

This video covers **BeautifulSoup** which is a python library that pulls data out of XML and HTML files. Documentation can be found on <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

The professor talked about why Dom and HTML became popular:

* Dom: The Document Object Model is an API that describes how webpages work, and it's implementation by browsers
* HTML: HTML text file outlining a page and a series of boxes representing the DOM nodes that match the elements

Useful link: https://mariechatfield.com/blog/html-vs-dom#:~:text=The%20DOM%20is%20the%20Document,nodes%20that%20match%20the%20elements.&text=We%20can%20see%20that%20the%20top%20level%20object%20is%20this%20document.

The hard part about web scrapping is that the layout of every web pages is different in structure.

Major emphasis was on how to use BeautifulSoup and then to covert it into the table:

* Naming the table
* What should be the fields in the table
* Absolute link: absolute URLs always include the domain name of the site with http://www
* Relative link: Relative links show the path to the file or refer to the file itself
* To check if it is a legitimate link—make it an absolute link—stick it into the database
* Date
* Check if data is scrapped twice (multiple ways this can be done depending on what of kind of data being used)

Useful web to check if the same url is being scrapped: https://stackoverflow.com/questions/3871613/scrapy-how-to-identify-already-scraped-urls

Using Python we can check if the web page is real or not hence we are going to use Python for web scrapping as opposed to just sql.

The next topic was on **how to install python mysql**:

1. Useful links: <https://www.youtube.com/watch?v=j6Pxrw5bnJU>
2. <https://www.youtube.com/watch?v=3vsC05rxZ8c>

The professor also mention about **Damen** which is a service that runs all the time