

Wee Kim Wee School of Communication and Information

**H6752-Data Extraction Techniques**

**Individual Report**

**Submitted By:**

**PHU WAI PAING (G1901561C)**

**Objective**

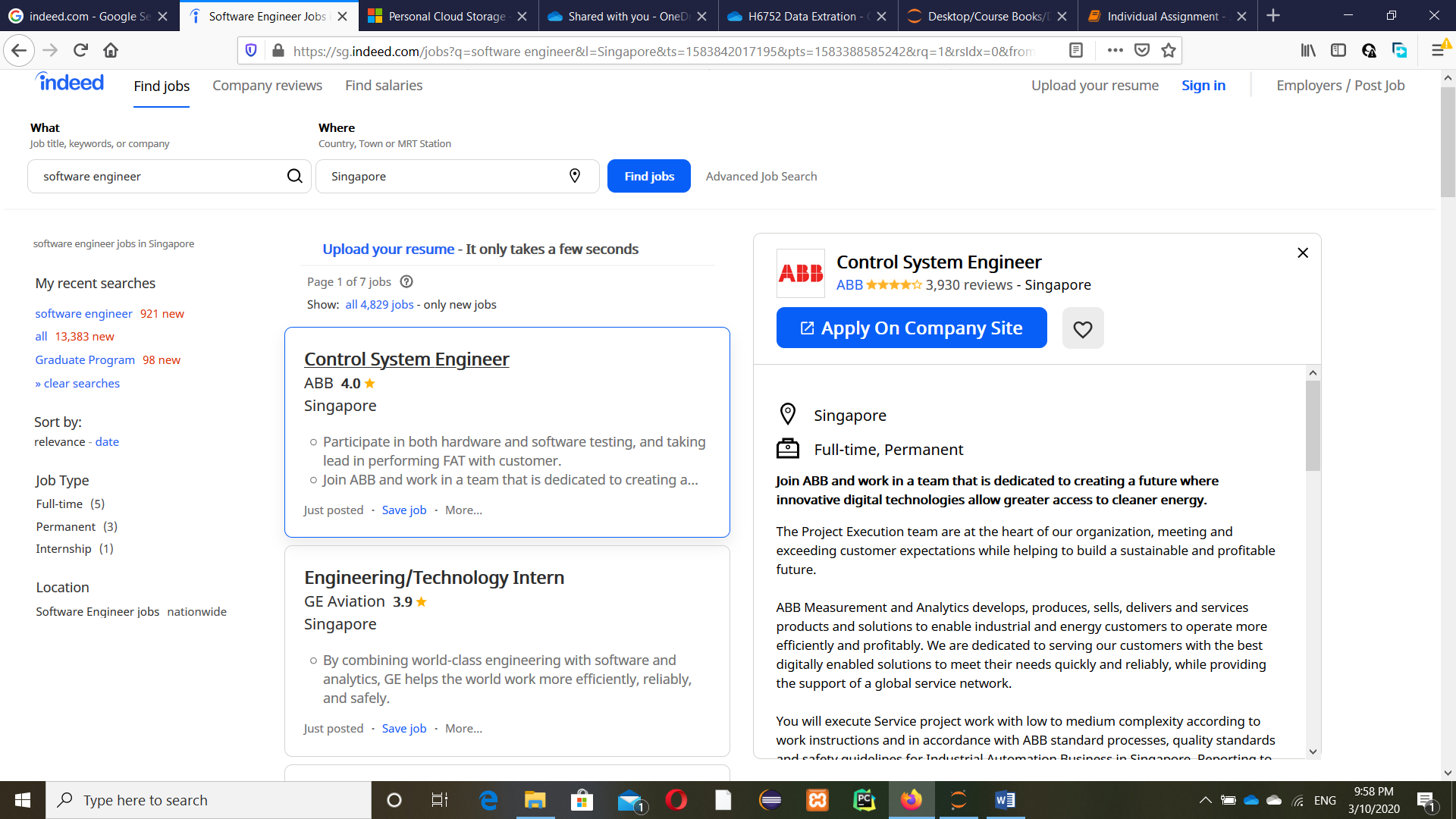
In this report, I will show how to extract job description information from indeed website using BeautifulSoup python library with selenium package. It helps to gather basic data regarding jobs posted on indeed. It can also help to monitor jobs that fit your profile, salary, and company and job title.

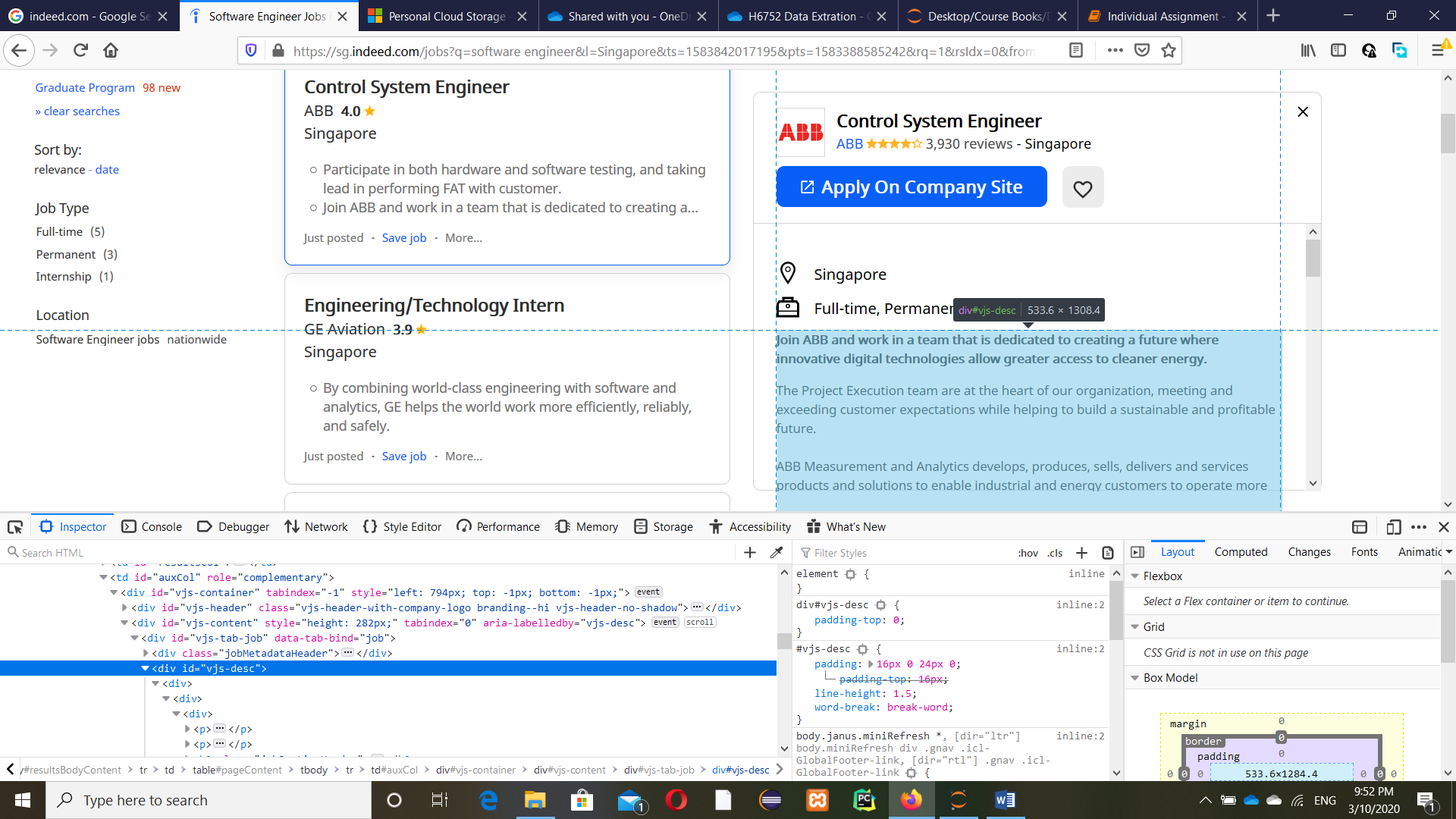
**Data Collection**

Firstly, I scraped the URLs from <https://sg.indeed.com/jobs?q=software+engineer&l=Singapore> which is started with jobs? Urls by using BeautifulSoup. The Url that I started to extract the links is related with software engineer position and Singapore place because I only want the data of software engineer position in Singapore. I saved these extracted urls in text file.

Next step is that collect the job description data from each link by clicking on an element when it has a class name to navigate the data on each link which I want by using Selenium package.

Below is an annotated screenshot of the data flow I would extract:





When you click Control System Engineer box, it will show the details information right hand side. So, I need to go each link to get the data I want. Then, scraped the job description with id representing the position of the data and saved each job description in each text file.

**Data Analysis**

After doing the data collection part, I would extract the salary from each job description data and analyze the text summarization for each one. I got the salary by using regular expression which started with ‘Salary’. If there were no matches, I replaced with ‘Negotiate’. For text summarization, there are five steps in generate summary.

Step 1: Read text of each line from text file

Step 2: Generate similarity matrix between sentences by using cosine similarity

Step 3: Rank sentences in similarity matrix with networkx python package

Step 4: Sort the rank and pick top sentences

Step 5: Output the summarize text and save in one text file

Finally, I saved all salary and text summarization data into one csv file again by using pandas library.

**Conclusion**

By extracting data and analyzing text summarization, we can save time on finding and reading information which we want. But, there is one limitation that if text file is too small, it doesn’t work properly in text summarization.