Midterm Project
ENGG 280 Winter 2022
Due: 8 March During the Lab

Project Logistics
It is recommended that you make a **group of a maximum of 5** for this project, however, you may also choose to work **individually**. All members in a group will receive the same grade.

Group Submissions: When submitting, ensure that all the **group members' names** are listed in the description. Only one group member needs to submit, but it is recommended that you complete the submission process together to double-check that nothing is missed! If the submitting group member uploads the wrong files, it will impact the grade of all members.

Program Specification

- **Stage1: Setup the project (Follow the steps to set up the project)**
  - Create a project called **ENGG680_Midterm**
  - Create 3 directories of
    - **src/**
      - Create a backend directory (**backend/**)
        - Inside the backend directory create **webscrap.py**
      - Create **main.py** file in src/ directory
    - **data/**
      - All results of running the code should be stored in the data directory.
    - **doc/**
      - Create a document including the name of the participant in the project.
        - **Members.pdf** (Fig1)

ENGG680 Midterm Project

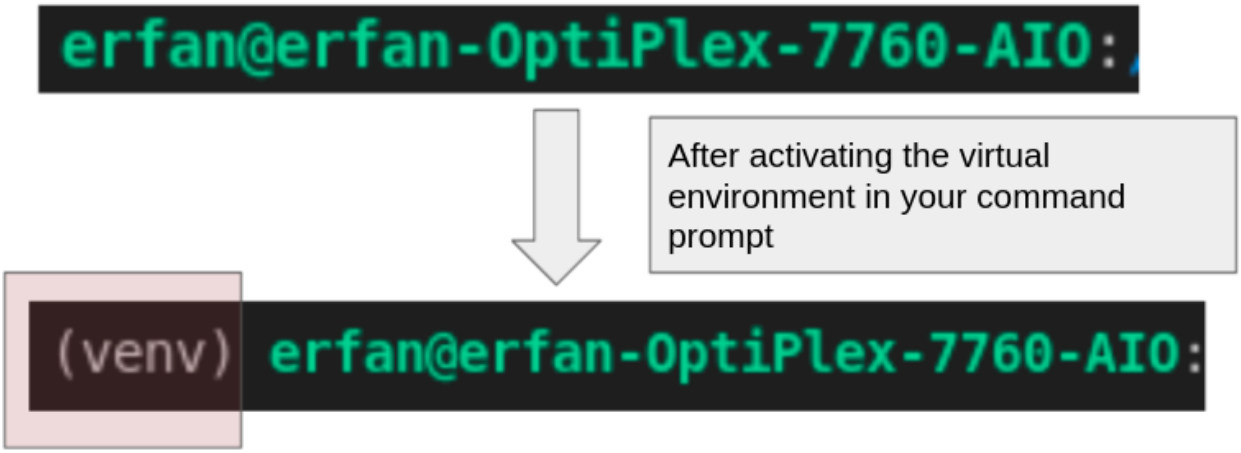
Member#1
- First Name: ..
- Last Name:..
- UCID:...

Member#2
- First Name: ..
- Last Name:..
- UCID:...

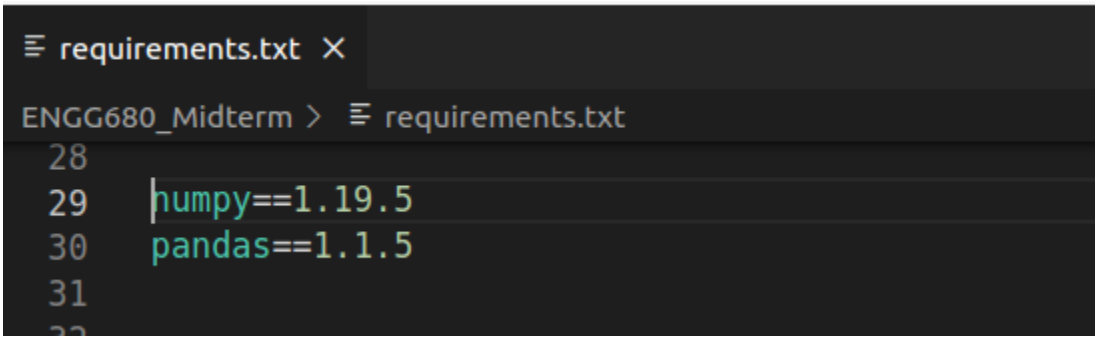
Fig1. The format of members.pdf file.

  - create a **venv/** directory inside ENGG680_Midterm by:
    - cmd$ python3.6 -m venv venv

- Activate your virtual environment and run all the codes by using this virtual environment:
  - cmd$ source /venv/bin/activate



- Install the following package in the virtual environment (venv/)
  - pip install beautifulsoup4
  - pip install requests
  - pip install pandas
  - pip install numpy
- Get log from the version of the packages that you are using and store it in the requirements.txt file:
  - pip freeze > **requirements.txt**



- In the end, your project must have all the directories are shown in Fig. 2 and Fig. 3.
- In Fig.3, you can see all the directories and the way they must be arranged in your project. **Note: Missing each part you may lose marks.**

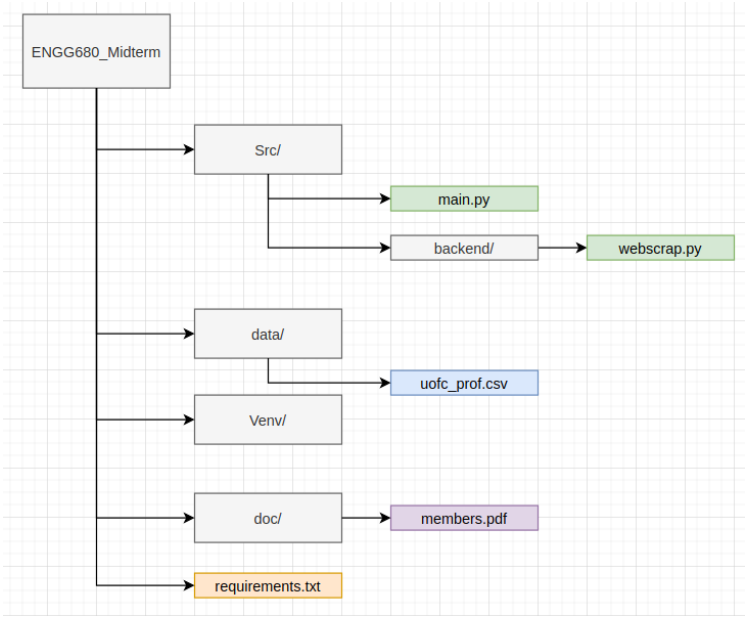


Fig. 2: The project files and directories Tree.

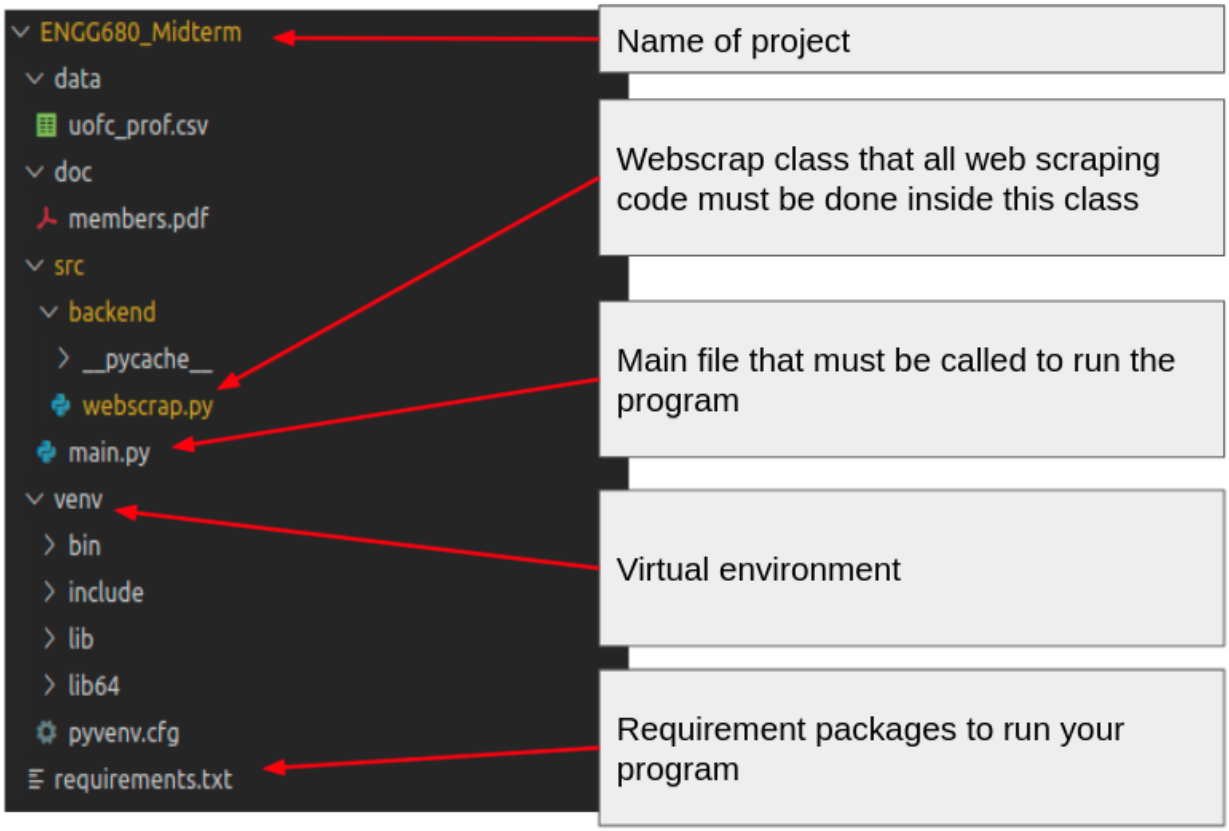


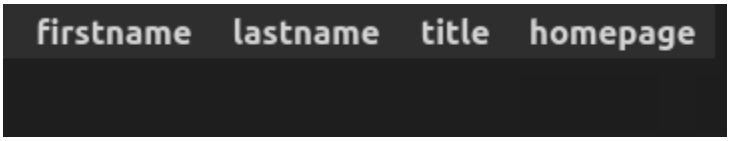
Fig. 3: Project Directories and Files and usage of each is explained.

- **Stage2:Crawl and Scrape**

- Schulich wants to have an integrated dataset of all Electrical and Engineering department professors in one place. So as a data engineer, you're asked to gather some information about engineering professors by crawling the faculty website. Then, scrape their information and load them to a pandas dataframe and eventually save it as a csv file.

  Website URL: https://schulich.ucalgary.ca/electrical-software/faculty-members

- In the first step, you need to get the html text of the website using `requests` library, and then you must use `Beautifulsoup4` library and `lxml` parser to parse the html and extract the needed information.
- Then, get the html text of the webpage and scrape the information of all its **Newest faculty members** and **professors** to put them in a dataframe as presented below:



- Tip: Use `Inspect Element` of Chrome to see the mapping html tags to objects in a webpage
  Tutorial Link: https://www.youtube.com/watch?v=1l4xz1QQhew&t=3s

- **Stage3: Explore the Data**
  - In this part, iterate on professors' dataframe and request to get their homepage html, and find the phone number and office (building and room) of each professor and add it to your previous dataframe as a new column. Finally, save the dataframe as a csv file in the data directory (**uofc_prof.csv**).

- **Stage4: Generating Report**
  - In this part, you need to generate the following reports:
    - Number of Assistant Professor
    - Number of Professor
    - Number of Senior Instructor
    - Number of Instructor
    - Number of Associate Professor