Pre-Work for Web Harvest Workshop

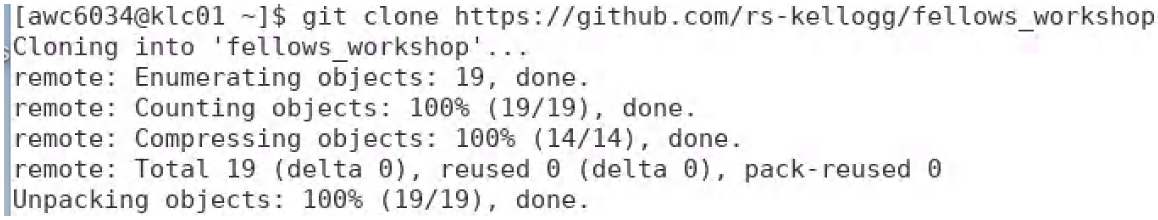
This workshop will reference a jupyter notebook for a web harvesting example. To use the notebook on KLC, please follow the steps below:

1.) Clone the github repository to your Home Directory

If you are not familiar with KLC, you can find an explanation and instructions for logging into the Linux Server here.

To clone the github repo, please sign in to KLC from FastX and follow the following steps:

- Open FastX from the web browser or your Desktop Application on any node
- Launch a GNOME Terminal window
- Type the following in the command line:
  `git clone https://github.com/rs-kellogg/fellows_workshop`

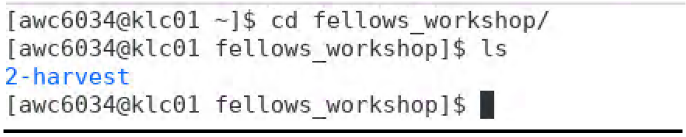


```
[awc6034@klc01 ~]$ git clone https://github.com/rs-kellogg/fellows_workshop
Cloning into 'fellows_workshop'...
remote: Enumerating objects: 19, done.
remote: Counting objects: 100% (19/19), done.
remote: Compressing objects: 100% (14/14), done.
remote: Total 19 (delta 0), reused 0 (delta 0), pack-reused 0
Unpacking objects: 100% (19/19), done.
```

2.) Update the github folder saved on KLC

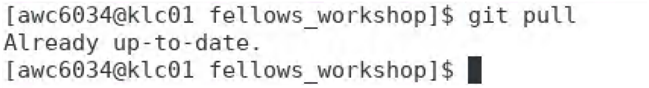
- To view the contents of the folder, type the following:

```
cd fellows-workshop
ls
```



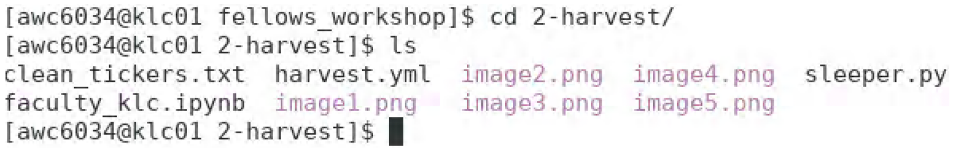
```
[awc6034@klc01 ~]$ cd fellows_workshop/
[awc6034@klc01 fellows_workshop]$ ls
2-harvest
[awc6034@klc01 fellows_workshop]$ 
```

- To update the folder you already downloaded, type:
  `git pull`



```
[awc6034@klc01 fellows_workshop]$ git pull
Already up-to-date.
[awc6034@klc01 fellows_workshop]$ 
```

- Change directories into **2-harvest by typing**
  `cd 2-harvest`



```
[awc6034@klc01 fellows_workshop]$ cd 2-harvest/
[awc6034@klc01 2-harvest]$ ls
clean_tickers.txt  harvest.yml  image2.png  image4.png  sleeper.py
faculty_klc.ipynb  image1.png   image3.png  image5.png
[awc6034@klc01 2-harvest]$ █
```

3.) Install web harvesting modules/packages in a conda environment

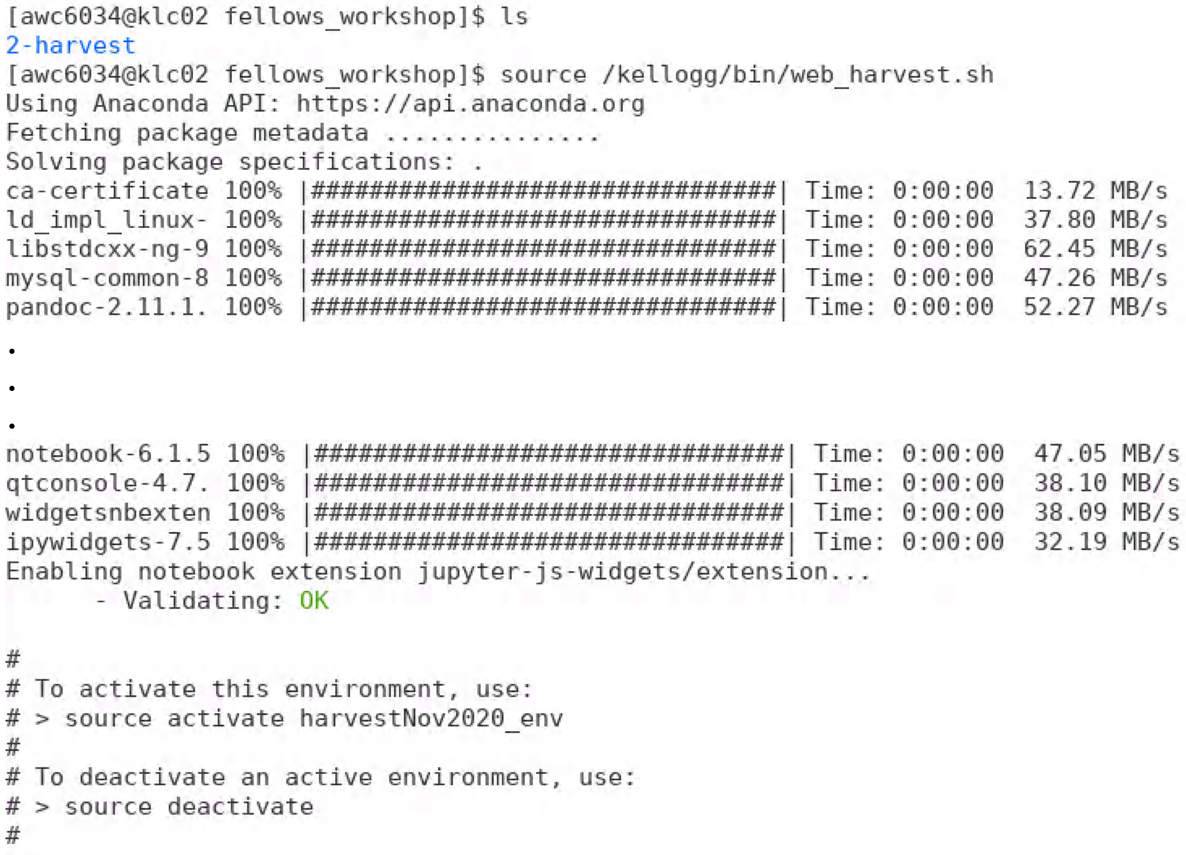
- Next, we will load python and the Firefox web browser. We will also create a conda environment (**harvestNov2020_env)** with the Beautiful Soup and selenium libraries installed. In order to complete this step, please make sure that **harvest.yml** is stored here:

  ```
  ~/empirical-workshop-2020/2-harvest
  ```

- Then type the following:

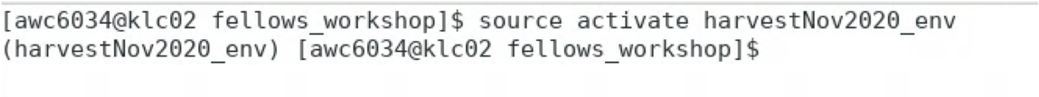
  ```
  source /kellogg/bin/web_harvest.sh
  ```

  It will take a while for the environment to install so please be patient.



```
[awc6034@klc02 fellows_workshop]$ ls
2-harvest
[awc6034@klc02 fellows_workshop]$ source /kellogg/bin/web_harvest.sh
Using Anaconda API: https://api.anaconda.org
Fetching package metadata ...............
Solving package specifications: .
ca-certificate 100% |##################################| Time: 0:00:00  13.72 MB/s
ld_impl_linux- 100% |##################################| Time: 0:00:00  37.80 MB/s
libstdcxx-ng-9 100% |##################################| Time: 0:00:00  62.45 MB/s
mysql-common-8 100% |##################################| Time: 0:00:00  47.26 MB/s
pandoc-2.11.1. 100% |##################################| Time: 0:00:00  52.27 MB/s
.
.
.
notebook-6.1.5 100% |##################################| Time: 0:00:00  47.05 MB/s
qtconsole-4.7. 100% |##################################| Time: 0:00:00  38.10 MB/s
widgetsnbexten 100% |##################################| Time: 0:00:00  38.09 MB/s
ipywidgets-7.5 100% |##################################| Time: 0:00:00  32.19 MB/s
Enabling notebook extension jupyter-js-widgets/extension...
      - Validating: OK

#
# To activate this environment, use:
# > source activate harvestNov2020_env
#
# To deactivate an active environment, use:
# > source deactivate
#
```

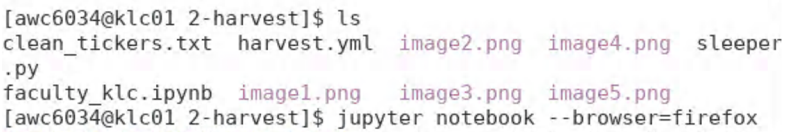
- Activate your conda environment by typing:
  `source activate harvestNov2020_env`



```
[awc6034@klc02 fellows_workshop]$ source activate harvestNov2020_env
(harvestNov2020_env) [awc6034@klc02 fellows_workshop]$
```

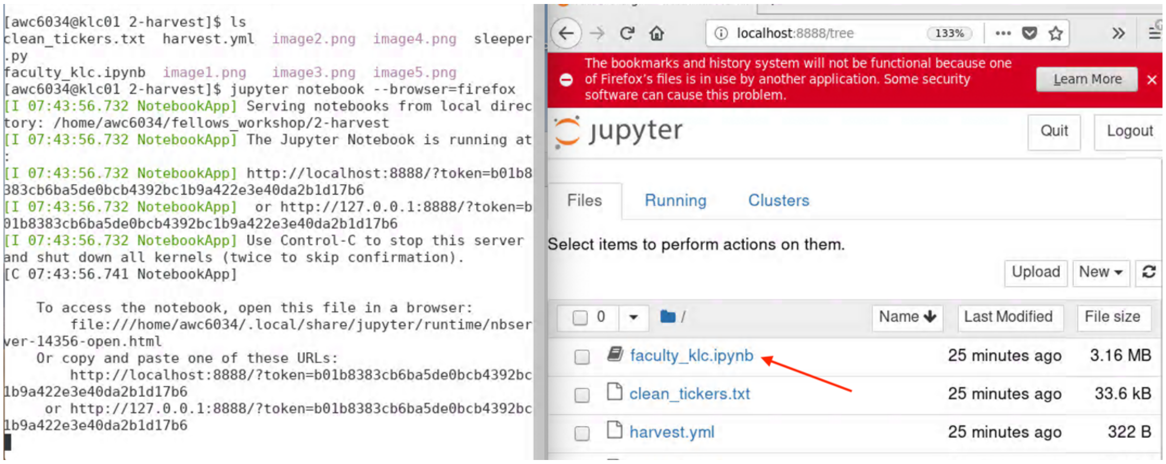
4.) Launch the jupyter notebook

- Launch the notebook by typing:
  `jupyter notebook --browser=firefox`



```
[awc6034@klc01 2-harvest]$ ls
clean_tickers.txt  harvest.yml  image2.png  image4.png  sleeper
.py
faculty_klc.ipynb  image1.png   image3.png  image5.png
[awc6034@klc01 2-harvest]$ jupyter notebook --browser=firefox
```

- A new Firefox window should launch. Just click on the Notebook. It is named faculty_klc.ipynb



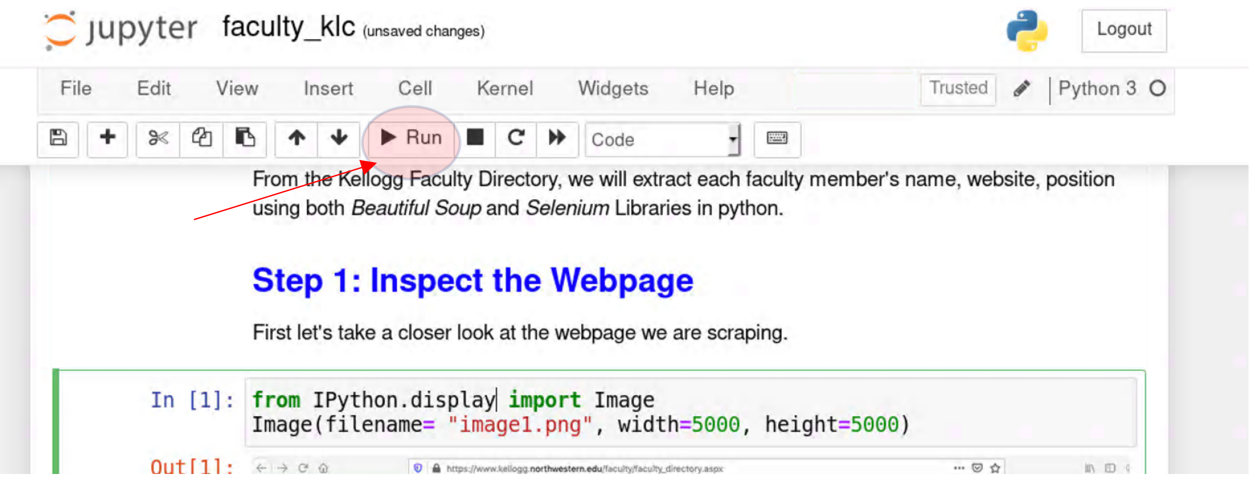
```
[awc6034@klc01 2-harvest]$ ls
clean_tickers.txt  harvest.yml  image2.png  image4.png  sleeper
.py
faculty_klc.ipynb  image1.png   image3.png  image5.png
[awc6034@klc01 2-harvest]$ jupyter notebook --browser=firefox
[I 07:43:56.732 NotebookApp] Serving notebooks from local direc
tory: /home/awc6034/fellows_workshop/2-harvest
[I 07:43:56.732 NotebookApp] The Jupyter Notebook is running at
:
[I 07:43:56.732 NotebookApp] http://localhost:8888/?token=b01b8
383cb6ba5de0bcb4392bc1b9a422e3e40da2b1d17b6
[I 07:43:56.732 NotebookApp]  or http://127.0.0.1:8888/?token=b
01b8383cb6ba5de0bcb4392bc1b9a422e3e40da2b1d17b6
[I 07:43:56.732 NotebookApp] Use Control-C to stop this server
and shut down all kernels (twice to skip confirmation).
[C 07:43:56.741 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///home/awc6034/.local/share/jupyter/runtime/nbser
ver-14356-open.html
    Or copy and paste one of these URLs:
        http://localhost:8888/?token=b01b8383cb6ba5de0bcb4392bc
1b9a422e3e40da2b1d17b6
     or http://127.0.0.1:8888/?token=b01b8383cb6ba5de0bcb4392bc
1b9a422e3e40da2b1d17b6
```

The bookmarks and history system will not be functional because one of Firefox's files is in use by another application. Some security software can cause this problem.

Select items to perform actions on them.

Name	Last Modified	File size
faculty_klc.ipynb	25 minutes ago	3.16 MB
clean_tickers.txt	25 minutes ago	33.6 kB
harvest.yml	25 minutes ago	322 B

- In the notebook, please confirm that you can run the code without errors by highlighting each line and clicking the RUN button



From the Kellogg Faculty Directory, we will extract each faculty member's name, website, position using both *Beautiful Soup* and *Selenium* Libraries in python.

Step 1: Inspect the Webpage

First let's take a closer look at the webpage we are scraping.

```
In [1]: from IPython.display import Image
        Image(filename= "image1.png", width=5000, height=5000)

Out[1]:
```

https://www.kellogg.northwestern.edu/faculty/faculty_directory.aspx

- When you are done with the notebook, press CTRL+C in the terminal window to stop it. Type `source deactivate harvestNov2020_env` to close the conda environment

- To activate the same conda environment after initially setting it up, type the following:

```
module load python/anaconda3.6
module load firefox/62
export PATH=/kellogg/bin:$PATH
export PYTHONNOUSERSITE=True
source activate harvestNov2020_env
```