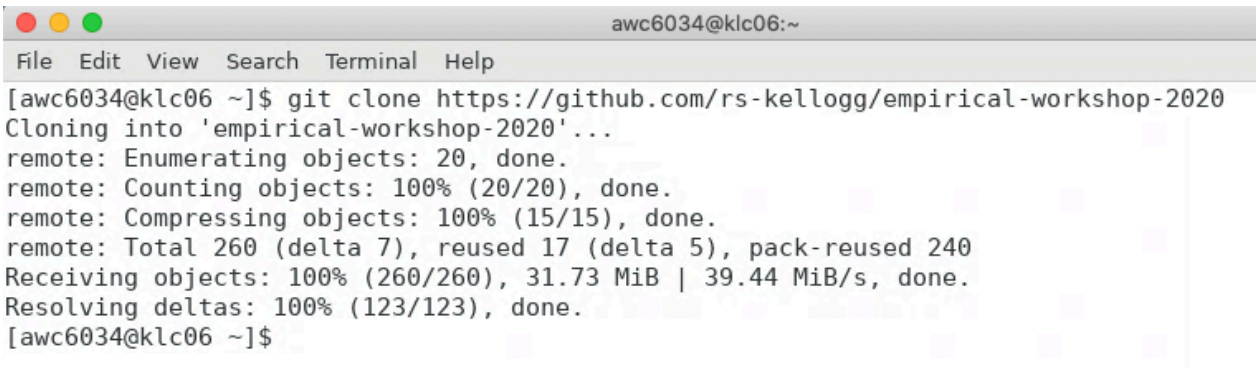
Pre-Work for Web Harvest Lecture

As always, you can link to a version of the jupyter notebook on AWS here: https://jupyterhub.rs-kellogg.org/hub/user-redirect/git-pull?repo=https%3A%2F%2Fgithub.com%2Frs-kellogg%2Fempirical-workshop-2020&subPath=5-harvest&app=lab

This week, we will use KLC to run our web harvesting examples. To use the notebook on KLC, please follow the steps below:

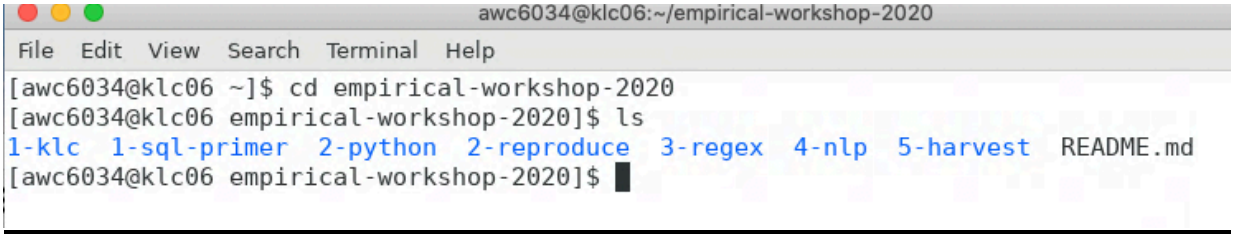
1.) Clone the github repository to your Home Directory

- 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/empirical-workshop-2020`



2.) Update the github folder saved on KLC

- To view the contents of the folder, type the following:
  `cd empirical-workshop-2020`
  `ls`



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

```
[awc6034@klc06 empirical-workshop-2020]$ git pull
Already up-to-date.
[awc6034@klc06 empirical-workshop-2020]$
```

- Change directories into **5-harvest by typing**
  `cd 5-harvest`

```
[awc6034@klc06 empirical-workshop-2020]$ ls
1-klc          2-python     3-regex   5-harvest
1-sql-primer   2-reproduce  4-nlp     README.md
[awc6034@klc06 empirical-workshop-2020]$ cd 5-harvest
[awc6034@klc06 5-harvest]$
```

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

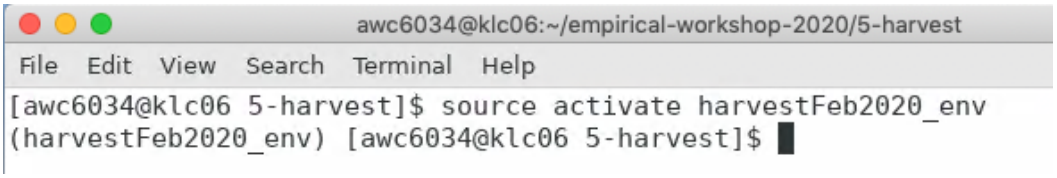
- Next, we will load python and the Firefox web browser. We will also create a conda environment (**harvestFeb2020_env)** with the Beautiful Soup and selenium libraries installed. In order to complete this step, please make sure that **harvest.yml** is stored here:

  `~/empirical-workshop-2020/5-harvest`

- Then type:
  `source /kellogg/bin/web_harvest.sh`

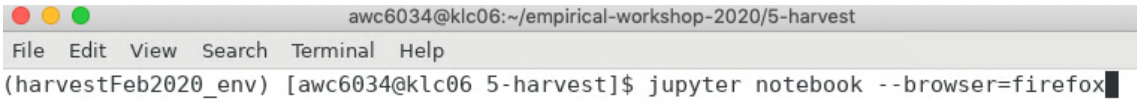
```
[awc6034@klc06 5-harvest]$ source /kellogg/bin/web_harvest.sh
Using Anaconda API: https://api.anaconda.org
Fetching package metadata ...............
Solving package specifications: .
Enabling notebook extension jupyter-js-widgets/extension...
      - Validating: OK

#
# To activate this environment, use:
# > source activate harvestFeb2020_env
#
# To deactivate an active environment, use:
# > source deactivate
#

[awc6034@klc06 5-harvest]$
```

- Activate your conda environment by typing:
  `source activate harvestFeb2020_env`

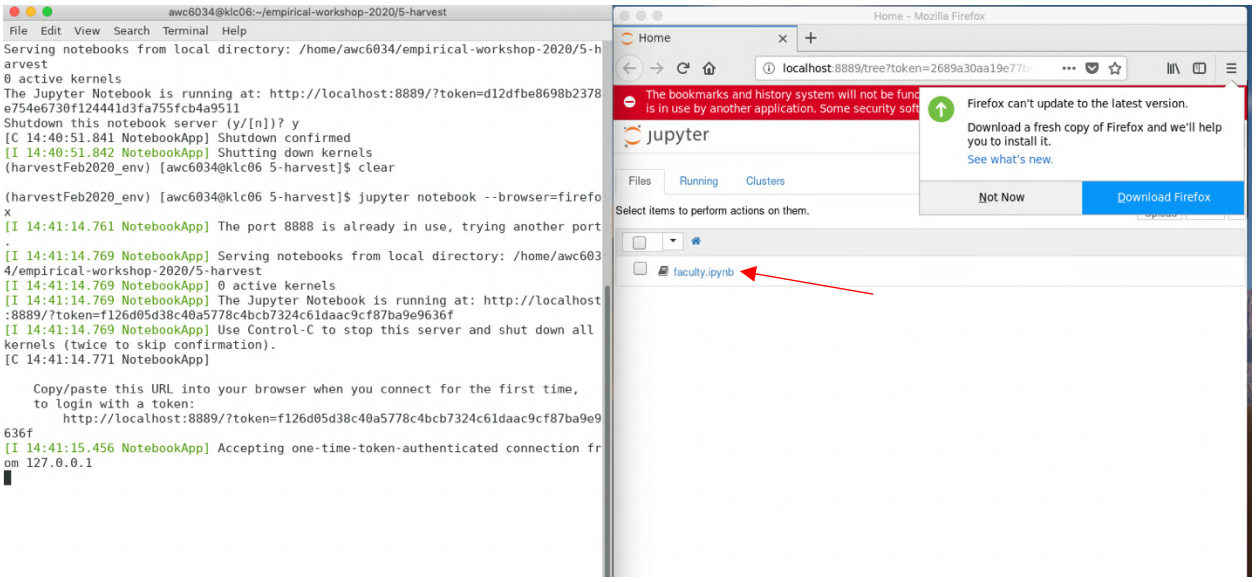


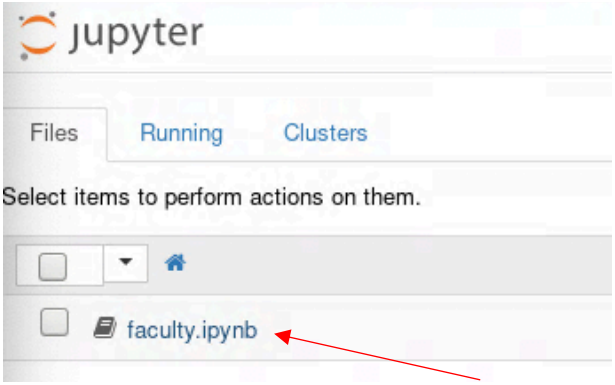
4.) Launch the jupyter notebook

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

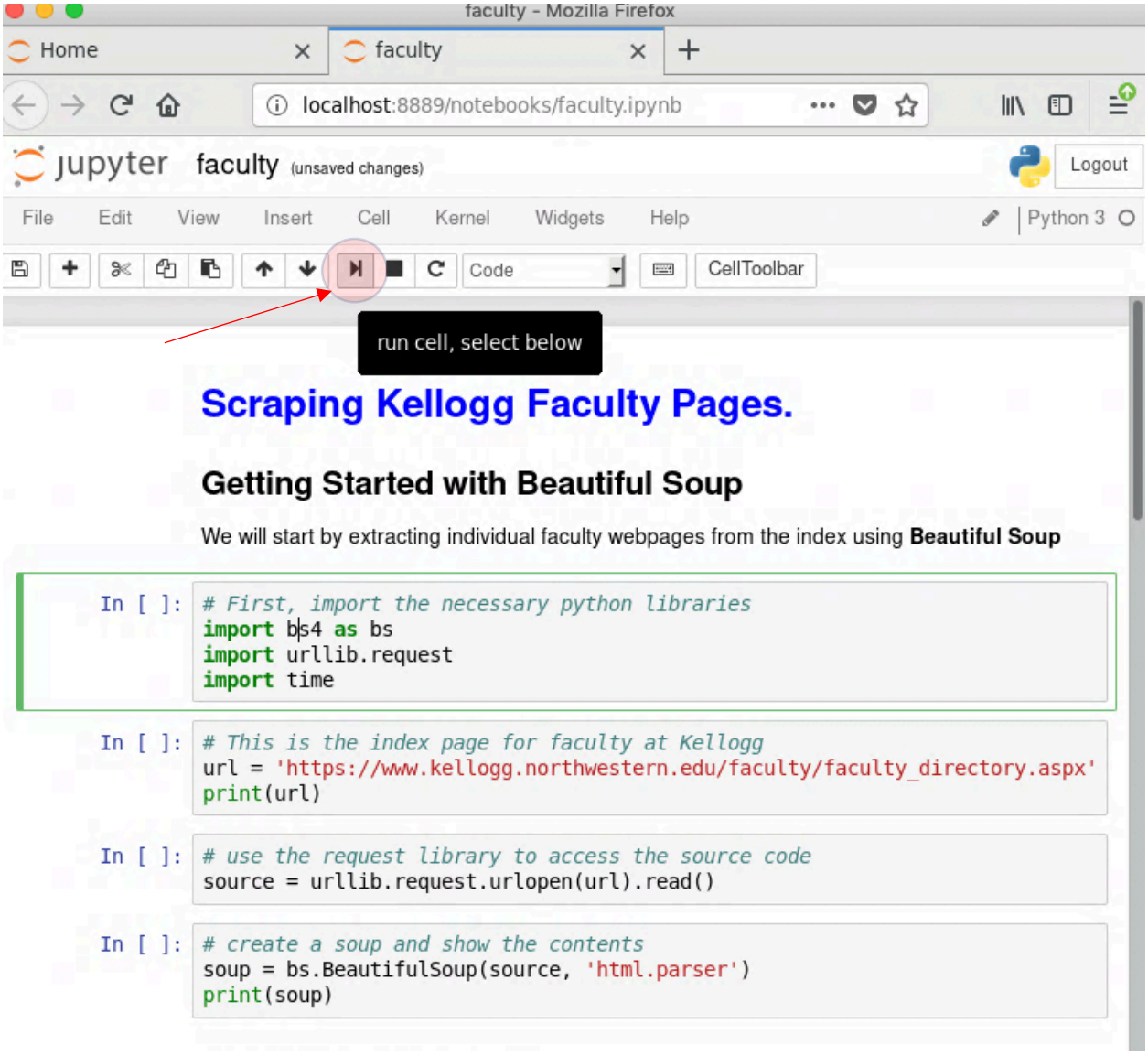


- A new Firefox window should launch. Just click on the Notebook





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



- When you are done with the notebook, press CTRL+C in the terminal window to stop it. Type `source deactivate harvestFeb2020_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 harvestFeb2020_env
```