Hw 7

In order to run my code on my local machine I had to install apache beam by using the command.

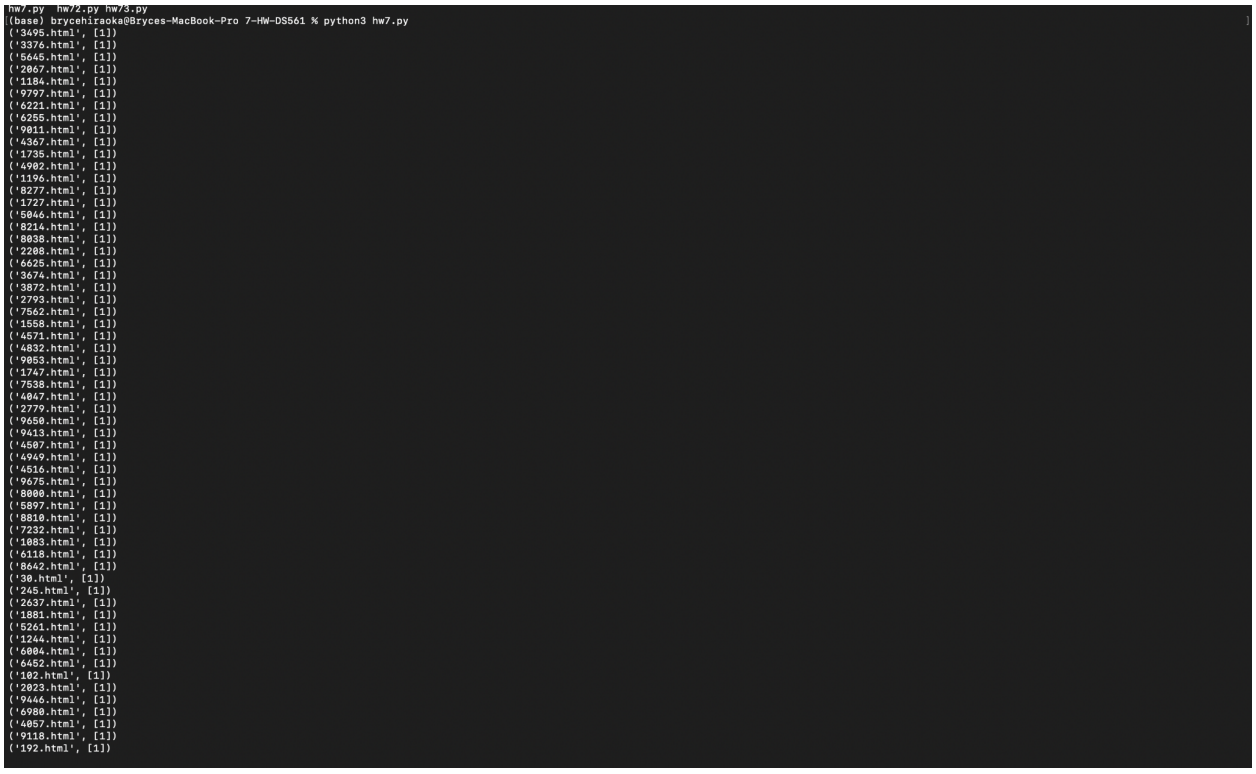
```
pip install 'apache-beam[gcp]'
```

To run the first application which finds and prints the files with the most incoming links I ran python3 hw7.py

I could not get the first application to work. No matter what I tried.

Whenever I tried to figure out the outgoing links all I got were a list of tuples.

Here is an example of a reading I got back.




Because I was unable to get it to work on my local machine I also could not get it to work on cloud dataflow.

To run the second application which finds and prints the files with the most outgoing links I ran python3 hw72.py

This took about an hour to run on my local machine and it returned

```
(base) brycehiraoka@crc-dot1x-nat-10-239-246-154 7-HW-DS561 % python3 hw72.py
[('2009.html', 14), ('8480.html', 13), ('5563.html', 13), ('3442.html', 13), ('4358.html', 12)]
(base) brycehiraoka@crc-dot1x-nat-10-239-246-154 7-HW-DS561 %
```

In order to run it in dataflow I had to add

```
google_cloud_options = options.view_as(GoogleCloudOptions)
   google_cloud_options.project = 'ds-561-398918'
   google_cloud_options.job_name = 'hw7_code'
   google_cloud_options.staging_location = 'gs://hw2-vm-bucket/webdir/staging'
   google_cloud_options.temp_location = 'gs://hw2-vm-bucket/webdir/temp'
   options.view_as(StandardOptions).runner = 'DataflowRunner'
```

To my python code.
I also had to make the output go to my cloud bucket instead of printing.

In order to run I used python3 hw73.py

It took about 15 min to run although I could not tell if it actually went through or not

Github: https://github.com/Bryce-Hiraoka/DS_561/tree/main/7-HW-DS561