

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.

```
hw7.py hw72.py hw73.py
(base) brycehiraoka@Bryces-MacBook-Pro 7-HW-DS561 % python3 hw7.py
('3495.html', [1])
('3376.html', [1])
('5645.html', [1])
('2867.html', [1])
('1184.html', [1])
('9797.html', [1])
('6221.html', [1])
('6255.html', [1])
('9811.html', [1])
('4367.html', [1])
('1735.html', [1])
('4982.html', [1])
('1196.html', [1])
('8277.html', [1])
('1727.html', [1])
('5846.html', [1])
('8214.html', [1])
('8838.html', [1])
('2288.html', [1])
('6625.html', [1])
('3674.html', [1])
('3872.html', [1])
('2793.html', [1])
('7562.html', [1])
('1558.html', [1])
('4571.html', [1])
('4832.html', [1])
('9853.html', [1])
('1747.html', [1])
('7538.html', [1])
('4847.html', [1])
('2779.html', [1])
('9658.html', [1])
('9413.html', [1])
('4587.html', [1])
('4949.html', [1])
('4516.html', [1])
('9675.html', [1])
('8888.html', [1])
('5897.html', [1])
('8818.html', [1])
('7232.html', [1])
('1883.html', [1])
('6118.html', [1])
('8642.html', [1])
('38.html', [1])
('245.html', [1])
('2637.html', [1])
('1881.html', [1])
('5261.html', [1])
('1244.html', [1])
('6886.html', [1])
('6452.html', [1])
('182.html', [1])
('2823.html', [1])
('9646.html', [1])
('4988.html', [1])
('4857.html', [1])
('9118.html', [1])
('192.html', [1])
```

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