DS561-HW7

U43188754

Github Link:

https://github.com/tigeryi1998/ds561-tigeryi/tree/main/hw7

git clone git@github.com:tigeryi1998/ds561-tigeryi.git

Install Apache Beam on both local and Google Cloud Shell

```
#!/bin/bash
pip3 install apache-beam
pip3 install 'apache-beam[gcp]'
pip3 install 'apache-beam[test]'
pip3 install 'apache-beam[docs]'

python3 hw7.py \
   --input gs://ds561-tigeryi-hw7/files/*.html \
   --output gs://ds561-tigeryi-hw7/output/ \
   --runner DataflowRunner \
   --project feisty-gasket-398719 \
   --region us-east1 \
   --temp_location gs://ds561-tigeryi-hw7/tmp/ \
   --staging_location gs://ds561-tigeryi-hw7/staging/ \
   --job_name job1

$ bash start.sh
```

Code for ParDo

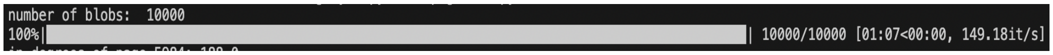
```
class ReadFileContent(beam.DoFn):
   def setup(self):
       self.storage_client = storage.Client()
   def process(self, filename):
       bucket_name="ds561-tigeryi-hw7"
       bucket = self.storage_client.get_bucket(bucket_name)
```

```
        blob = bucket.get_blob(filename)
        content = blob.download_as_string().decode("utf-8")
        links = re.findall(r'<a HREF="(\d+).html">', content)
        links_int = [int(x) for x in links]
        filename_int = int(blob.name.split(".")[0].split("/")[1])
        yield (filename_int, links_int)
```

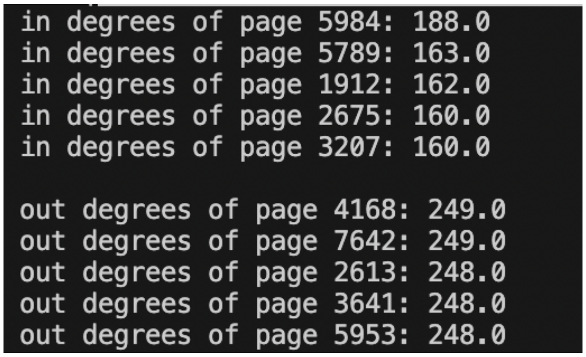
The will return the page name, and a list of all pages that it points to.

Then can build the out-matrix and in-matrix. (10000 x 10000)

Result:



```
number of blobs:  10000
100%|████████████████████████████████████████| 10000/10000 [01:07<00:00, 149.18it/s]
```



```
in degrees of page 5984: 188.0
in degrees of page 5789: 163.0
in degrees of page 1912: 162.0
in degrees of page 2675: 160.0
in degrees of page 3207: 160.0

out degrees of page 4168: 249.0
out degrees of page 7642: 249.0
out degrees of page 2613: 248.0
out degrees of page 3641: 248.0
out degrees of page 5953: 248.0
```

top 5 files with the most incoming links

in degrees of page 5984: 188.0

in degrees of page 5789: 163.0

in degrees of page 1912: 162.0

in degrees of page 2675: 160.0

in degrees of page 3207: 160.0

top 5 files with the most Outgoing links

```
out degrees of page 4168: 249.0

out degrees of page 7642: 249.0

out degrees of page 2613: 248.0

out degrees of page 3641: 248.0

out degrees of page 5953: 248.0
```

It takes 10-15 mins on the GCP and 50-55 mins on the local