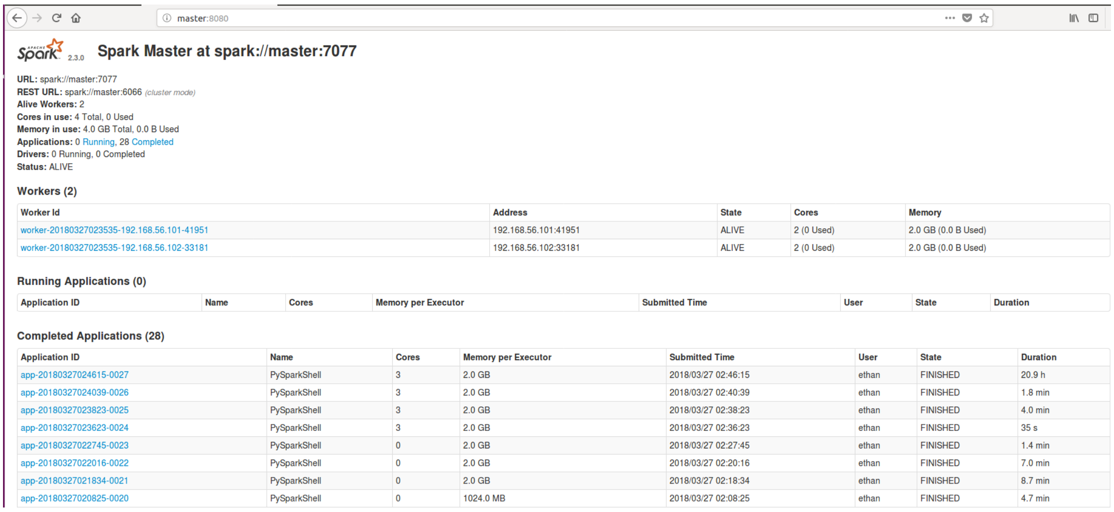
Big Data Mining and Applications

Homework3
106598005 盧家馨
2018/05/04

- My cluster environment setup:

PC	Memory	Ip address
Master	2G	192.168.56.100
Node1	2G	192.168.56.101
Node2	2G	192.168.56.102

- Connection Status



Spark 2.3.0 Spark Master at spark://master:7077

URL: spark://master:7077
REST URL: spark://master:6066 *(cluster mode)*
Alive Workers: 2
Cores in use: 4 Total, 0 Used
Memory in use: 4.0 GB Total, 0.0 B Used
Applications: 0 Running, 28 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory
worker-20180327023535-192.168.56.101-41951	192.168.56.101:41951	ALIVE	2 (0 Used)	2.0 GB (0.0 B Used)
worker-20180327023535-192.168.56.102-33181	192.168.56.102:33181	ALIVE	2 (0 Used)	2.0 GB (0.0 B Used)

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

Completed Applications (28)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20180327024615-0027	PySparkShell	3	2.0 GB	2018/03/27 02:46:15	ethan	FINISHED	20.9 h
app-20180327024039-0026	PySparkShell	3	2.0 GB	2018/03/27 02:40:39	ethan	FINISHED	1.8 min
app-20180327023823-0025	PySparkShell	3	2.0 GB	2018/03/27 02:38:23	ethan	FINISHED	4.0 min
app-20180327023623-0024	PySparkShell	3	2.0 GB	2018/03/27 02:36:23	ethan	FINISHED	35 s
app-20180327022745-0023	PySparkShell	0	2.0 GB	2018/03/27 02:27:45	ethan	FINISHED	1.4 min
app-20180327022016-0022	PySparkShell	0	2.0 GB	2018/03/27 02:20:16	ethan	FINISHED	7.0 min
app-20180327021834-0021	PySparkShell	0	2.0 GB	2018/03/27 02:18:34	ethan	FINISHED	8.7 min
app-20180327020825-0020	PySparkShell	0	1024.0 MB	2018/03/27 02:08:25	ethan	FINISHED	4.7 min

- Source code link
  https://github.com/Jessieluu/Spark2018/tree/master/hw3

- Output link
  https://drive.google.com/drive/folders/1rS-mA7p2B_-eQrIPIip6cya7yKUvJ-WW?usp=sharing

Because the file is too large to compute, so I choose 5 document to compute and generate the matrix with (shingle * document(5))