

Big Data Mining and Applications - HW3

Lars Fikkers (108012050), Pascal Roose (108012051)

About

The homework assignments were made with PySpark and executed/tested on Jupyter Notebook. Feel free to login and look around and execute the homework assignments.

Webserver: <https://bdm.pjepos.nl/>

Password: 44p9uj@93ArD

You will find the source files in the folder src/ and the output in out/

You can also check out our git repository on GitHub which includes all input, output and source files. <https://github.com/PascalRoose/bigdatamining/>
We didn't include output this time because of the actual size of the output and memory issues we had when finishing the homework.

Detailed Responsibility

Member	Part
Pascal Roose	Exercise 1 & (part of) 2
Lars Fikkers	(part of) Exercise 2

Environment setup

Host: Digital Ocean

(we changed the amount of memory and virtual CPUs we had multiple times to speed things up)

Specs:

- Ubuntu 18.04.3 (LTS) x64
- 2 vCPUs
- 4GB RAM / 25GB ROM
- Region Singapore(l)

Spark:

- Cluster mode
 - Master: <http://104.248.150.122:8080/>
 - Worker: <http://104.248.150.122:8081/>

- With Python (PySpark)
- With Jupyter Notebook as driver

Setup and installation

Requirements

```
sudo apt install python3  
sudo apt install pip-python3
```

Apache Spark

Download, unpack and move

```
wget https://www-eu.apache.org/dist/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz  
tar xvf spark-2.4.4-bin-hadoop2.7.tgz  
mv spark-2.4.4-bin-hadoop2.7 /usr/local/spark
```

Create then edit the configuration file using nano

```
mv /usr/local/spark/conf/  
cp spark-env.sh.template spark-env.sh  
sudo nano spark-env.sh
```

Add the following line: SPARK_MASTER_HOST = 'your ip-address'

Exit using Ctrl+x, press enter to save

Start the master- and worker-node

```
cd /usr/local/spark/sbin/  
./start-master.sh  
./start-slave.sh spark://ip-address:port
```

PySpark

```
pip3 install pyspark
```

Jupyter Notebook

```
pip3 install jupyter
```

Connect Jupyter Notebook to Spark

Setup environment variables

```
export PATH=$PATH:/usr/local/spark/bin  
export SPARK_HOME=/usr/local/spark  
export PYSARK_DRIVER_PYTHON=jupyter  
export PYSARK_DRIVER_PYTHON_OPTS='notebook'
```

Run Jupyter Notebook with PySpark

```
cd /path/to/homework/src  
pyspark
```

