Big Data Mining and Applications - HW4

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/

Detailed Responsibility

Member	Part
Pascal Roose	
Lars Fikkers	

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
- 4 vCPUs
- 16 GB RAM / 25 GB ROM
- Region Singapore(1)

Spark:

- 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 PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'
```

Run Jupyter Notebook with PySpark

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