

Running review.py:

Folder Structure:

I have placed all my files under the directory /home/dk/Desktop/pythonhw/hw3/*

So kindly change the path accordingly inside the program review.py.

Sparkhw →

Review.py → main file

Review.data, meta.data → input files

Output1 Output2 → output folders

Execute the following:

Move to spark folder and then to sbin folder

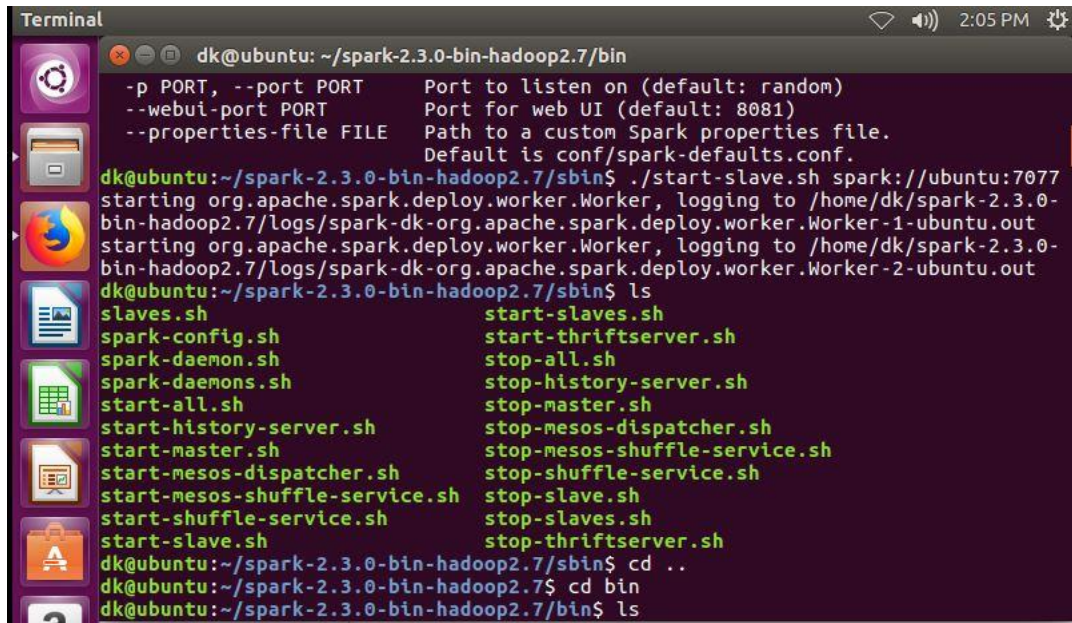
./start-master.sh

```
dk@ubuntu: ~/spark-2.3.0-bin-hadoop2.7/bin
dk@ubuntu:~$ cd ~/spark-2.3.0-bin-hadoop2.7/sbin
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/sbin$ ./start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /home/dk/spark-2.3.0-
bin-hadoop2.7/logs/spark-dk-org.apache.spark.deploy.master.Master-1-ubuntu.out
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/sbin$ cd ..
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7$ ls
bin  data  jars  LICENSE  logs  python  README.md  sbin  yarn
conf  examples  kubernetes  licenses  NOTICE  R  RELEASE  work
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7$ cd conf
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/conf$ ls
docker.properties.template  slaves.template
fairscheduler.xml.template  spark-defaults.conf.template
log4j.properties           spark-env.sh
log4j.properties.template  spark-env.sh.template
metrics.properties.template
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/conf$ cd sbin
bash: cd: sbin: No such file or directory
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/conf$ cd ..
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7$ cd sbin
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/sbin$ ./start-slave.sh
Usage: ./sbin/start-slave.sh [options] <master>

Master must be a URL of the form spark://hostname:port
```

Start slave with master url:

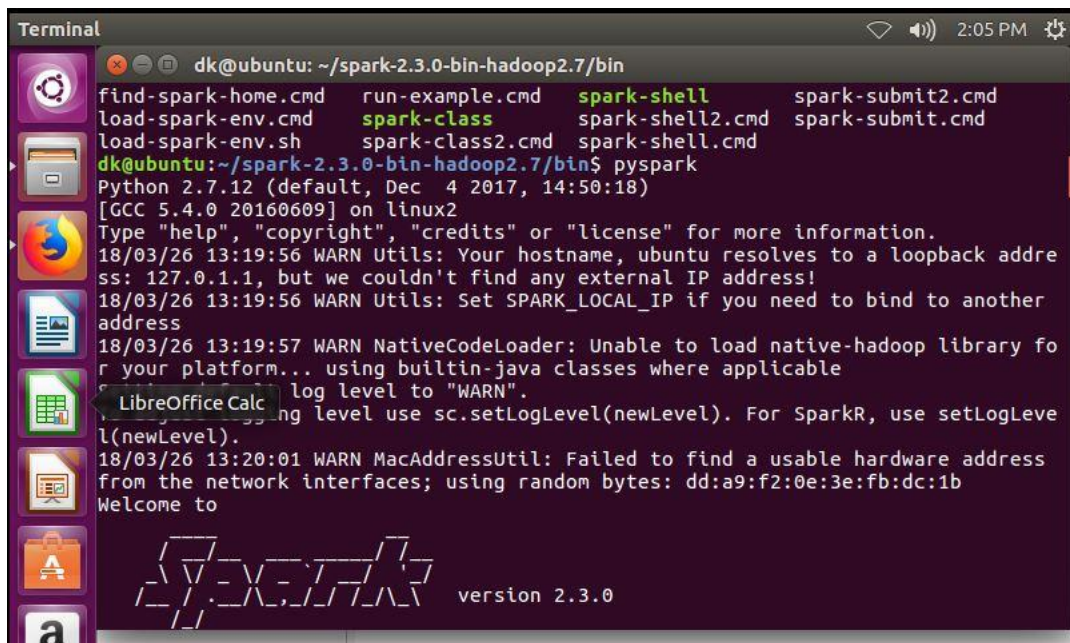
./start-slave.sh spark://Ubuntu:7077

A terminal window titled 'Terminal' with a dark background and light text. The prompt is 'dk@ubuntu: ~/spark-2.3.0-bin-hadoop2.7/bin'. The user enters './start-slave.sh spark://ubuntu:7077'. The output shows the command-line options for the script: '-p PORT, --port PORT Port to listen on (default: random)', '--webui-port PORT Port for web UI (default: 8081)', and '--properties-file FILE Path to a custom Spark properties file. Default is conf/spark-defaults.conf.' This is followed by two log messages: 'starting org.apache.spark.deploy.worker.Worker, logging to /home/dk/spark-2.3.0-bin-hadoop2.7/logs/spark-dk-org.apache.spark.deploy.worker.Worker-1-ubuntu.out' and 'starting org.apache.spark.deploy.worker.Worker, logging to /home/dk/spark-2.3.0-bin-hadoop2.7/logs/spark-dk-org.apache.spark.deploy.worker.Worker-2-ubuntu.out'. Then the user enters 'ls' and a list of files is shown: 'slaves.sh', 'spark-config.sh', 'spark-daemon.sh', 'spark-daemons.sh', 'start-all.sh', 'start-history-server.sh', 'start-master.sh', 'start-mesos-dispatcher.sh', 'start-mesos-shuffle-service.sh', 'start-shuffle-service.sh', 'start-slave.sh', 'start-slaves.sh', 'start-thriftserver.sh', 'stop-all.sh', 'stop-history-server.sh', 'stop-master.sh', 'stop-mesos-dispatcher.sh', 'stop-mesos-shuffle-service.sh', 'stop-shuffle-service.sh', 'stop-slave.sh', 'stop-slaves.sh', and 'stop-thriftserver.sh'. Finally, the user enters 'cd ..' and 'cd bin', and the prompt changes to 'dk@ubuntu: ~/spark-2.3.0-bin-hadoop2.7/bin\$'.

```
Terminal
dk@ubuntu: ~/spark-2.3.0-bin-hadoop2.7/bin
-p PORT, --port PORT      Port to listen on (default: random)
--webui-port PORT         Port for web UI (default: 8081)
--properties-file FILE    Path to a custom Spark properties file.
                          Default is conf/spark-defaults.conf.
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/sbin$ ./start-slave.sh spark://ubuntu:7077
starting org.apache.spark.deploy.worker.Worker, logging to /home/dk/spark-2.3.0-
bin-hadoop2.7/logs/spark-dk-org.apache.spark.deploy.worker.Worker-1-ubuntu.out
starting org.apache.spark.deploy.worker.Worker, logging to /home/dk/spark-2.3.0-
bin-hadoop2.7/logs/spark-dk-org.apache.spark.deploy.worker.Worker-2-ubuntu.out
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/sbin$ ls
slaves.sh                start-slaves.sh
spark-config.sh          start-thriftserver.sh
spark-daemon.sh          stop-all.sh
spark-daemons.sh        stop-history-server.sh
start-all.sh            stop-master.sh
start-history-server.sh  stop-mesos-dispatcher.sh
start-master.sh          stop-mesos-shuffle-service.sh
start-mesos-dispatcher.sh stop-shuffle-service.sh
start-mesos-shuffle-service.sh stop-slave.sh
start-shuffle-service.sh stop-slaves.sh
start-slave.sh           stop-thriftserver.sh
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/sbin$ cd ..
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7$ cd bin
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/bin$ ls
```

Check localhost for active clusters:

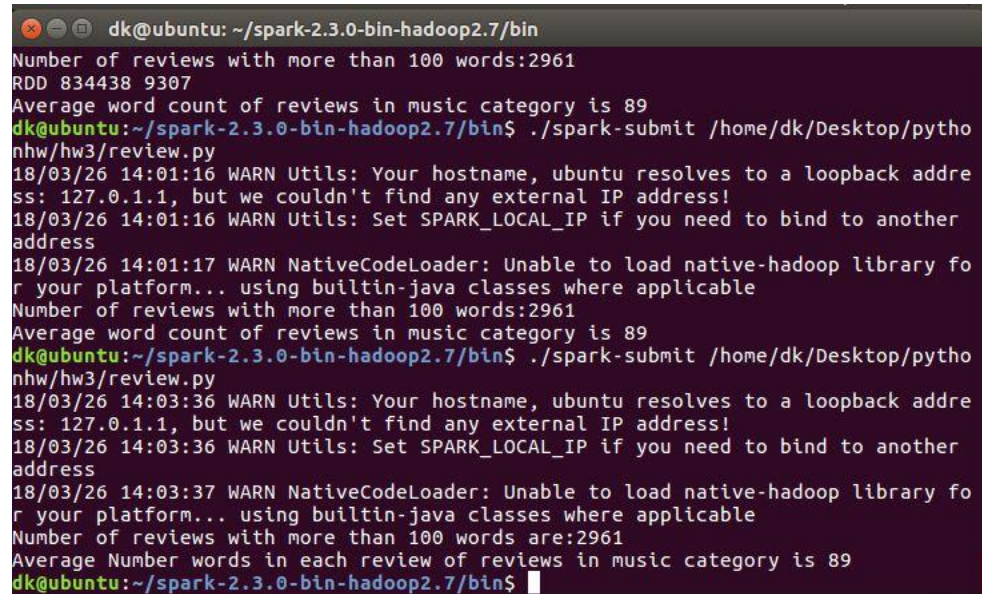
https://localhost:8088



Use sparksubmit under bin folder to execute review.py:

`./spark-submit /home/dk/Desktop/pythonhw/hw3/review.py`

You can find the output printed.

A terminal window with a dark background and light text. The title bar shows 'dk@ubuntu: ~/spark-2.3.0-bin-hadoop2.7/bin'. The output of the command shows the results of a Spark job. It includes the number of reviews with more than 100 words (2961), the RDD (834438 9307), and the average word count of reviews in the music category (89). There are also several warning messages from the Spark Utils and NativeCodeLoader. The command is repeated twice, and the output is identical each time.

```
dk@ubuntu: ~/spark-2.3.0-bin-hadoop2.7/bin
Number of reviews with more than 100 words:2961
RDD 834438 9307
Average word count of reviews in music category is 89
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/bin$ ./spark-submit /home/dk/Desktop/pythonhw/hw3/review.py
18/03/26 14:01:16 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1, but we couldn't find any external IP address!
18/03/26 14:01:16 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
18/03/26 14:01:17 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Number of reviews with more than 100 words:2961
Average word count of reviews in music category is 89
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/bin$ ./spark-submit /home/dk/Desktop/pythonhw/hw3/review.py
18/03/26 14:03:36 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1, but we couldn't find any external IP address!
18/03/26 14:03:36 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
18/03/26 14:03:37 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Number of reviews with more than 100 words are:2961
Average Number words in each review of reviews in music category is 89
dk@ubuntu:~/spark-2.3.0-bin-hadoop2.7/bin$
```