CS5344 Lab 1

AY2021/2022 Semester 2

The purpose of this lab is to get you started with Spark, and learn how to write, compile, debug and execute a simple Spark program. You will also be tasked to write and submit your own Spark program individually.

1. **Reference programs and documentation** from Spark release are available at https://www.tutorialspoint.com/apache_spark/apache_spark_quick_guide.htm https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html

2. A VirtualBox image of Ubuntu with Spark deployment has been configured for you. **Appendix A** gives the instructions on how to download and install it.

3. Alternatively, you can learn to install a stand-alone Spark-2.2.1 instance on Ubuntu by yourself and set up the environment by following the instructions in **Appendix B.**
   For **Mac users,** you can refer to
   https://medium.com/luckspark/installing-spark-2-3-0-on-macos-high-sierra-276a127b8b85
   https://notadatascientist.com/install-spark-on-macos/

4. **Appendix C** is the basic guide to help you get started and run your first Spark *WordCount* program in Python 3.6.

5. If you want to **debug with PyCharm**, you can link PyCharm with Spark according to the instruction in https://stackoverflow.com/questions/34685905/how-to-link-pycharm-with-pyspark.

   To install PyCharm and run your first project, you can refer to https://www.jetbrains.com/help/pycharm/installation-guide.html

Task: Write a Spark program to find the top 15 products based on the number of user reviews and report their average rating and product price.

Datasets:

Use the Baby review file (reviews_Baby. json) and metadata (meta_Baby.json) from the Amazon product dataset (http://jmcauley.ucsd.edu/data/amazon/links.html).
Download both files from the "Per-category files" section.

Algorithm:

Step 1. Find the number of reviews and calculate the average rating for each product from the review file. Use pair RDD, reduceByKey and map function to accomplish this step. The key is the product ID/asin.

Step 2. Create an RDD where the key is the product ID/asin and value is the price of the product. Use the metadata for this step.

Step 3. Join the pair RDD obtained in Step 1 and the RDD created in Step 2.

Step 4. Find the top 15 products with the greatest number of reviews.

Step 5. Output the average rating and price for the top 15 products identified in Step 4.

Input: Review file and metadata.

Output: One line per product in the following format:

<product ID> <average rating> <product price>

Deliverables:

Zip the Spark program with documentation for important steps in the code along with the output file and upload it to Lab1 folder. The zipped folder should be named as follows, StudentID_Lab1.

Important Notes:

(a) Specify the python version and the packages used.
(b) Your code should be executable either on the given virtual machine configuration given or on stand-alone Spark configuration.

References:

- https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html#transformations
- https://spark.apache.org/examples.html

Appendix A. Install VirtualBox and Configure Ubuntu Image

1. Install VirtualBox VM https://www.virtualbox.org/

   VirtualBox supports Windows, Mac OS, Linux, Solaris.

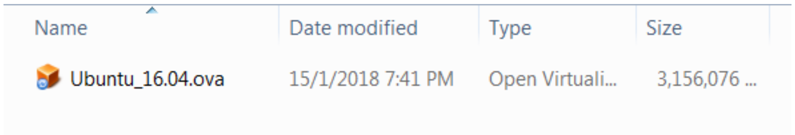
   Download the installation file for your operating system. Double-click on file to install it.

2. Download Ubuntu Image from

   https://nusu-my.sharepoint.com/:u:/g/personal/e0321289_u_nus_edu/EaLQ2hD4uotCkmAY9n7uWyMB6E6PlCbntXXvywHJBh5AKw

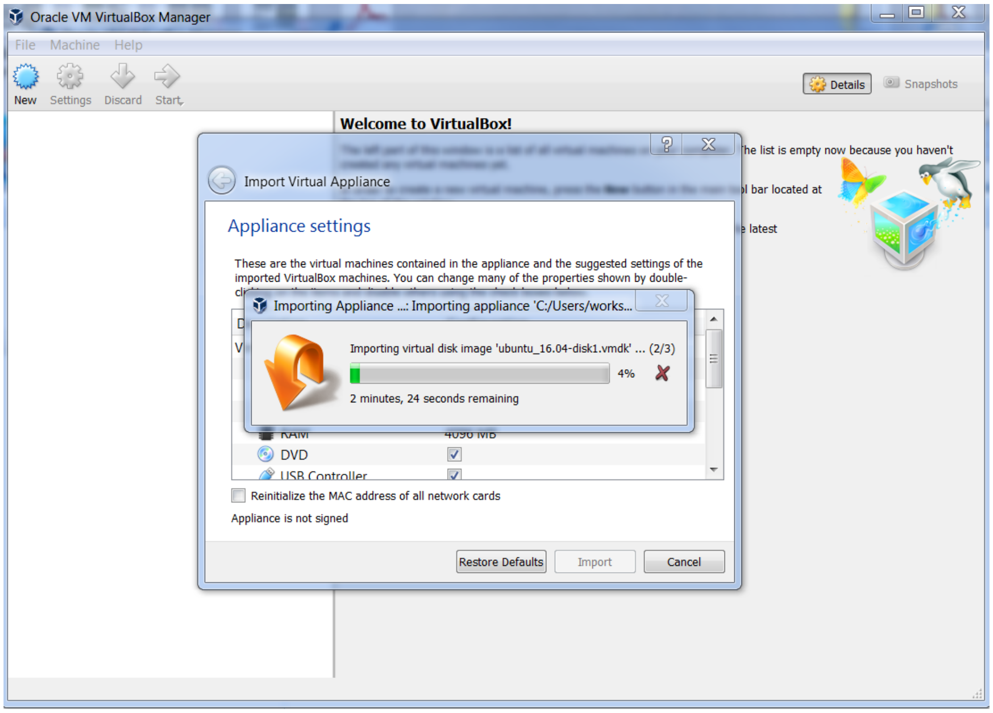
   When prompted to sign-in use your nus mail id (xxx@u.nus.edu), this should take you to the nus mail page. Log-in to download the zip file.

   You will see the following file when you unzip ubuntu_16.04.zip file:



Name	Date modified	Type	Size
Ubuntu_16.04.ova	15/1/2018 7:41 PM	Open Virtuali...	3,156,076 ...

3. Double click Ubuntu_16.04.vbox file, and the image will be loaded into VirtualBox.



Oracle VM VirtualBox Manager

Welcome to VirtualBox!

he list is empty now because you haven't

l bar located at

latest

Import Virtual Appliance

Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-

Importing Appliance ...: Importing appliance 'C:/Users/works...

Importing virtual disk image 'ubuntu_16.04-disk1.vmdk' ... (2/3)

4%

2 minutes, 24 seconds remaining

RAM 4096 MB

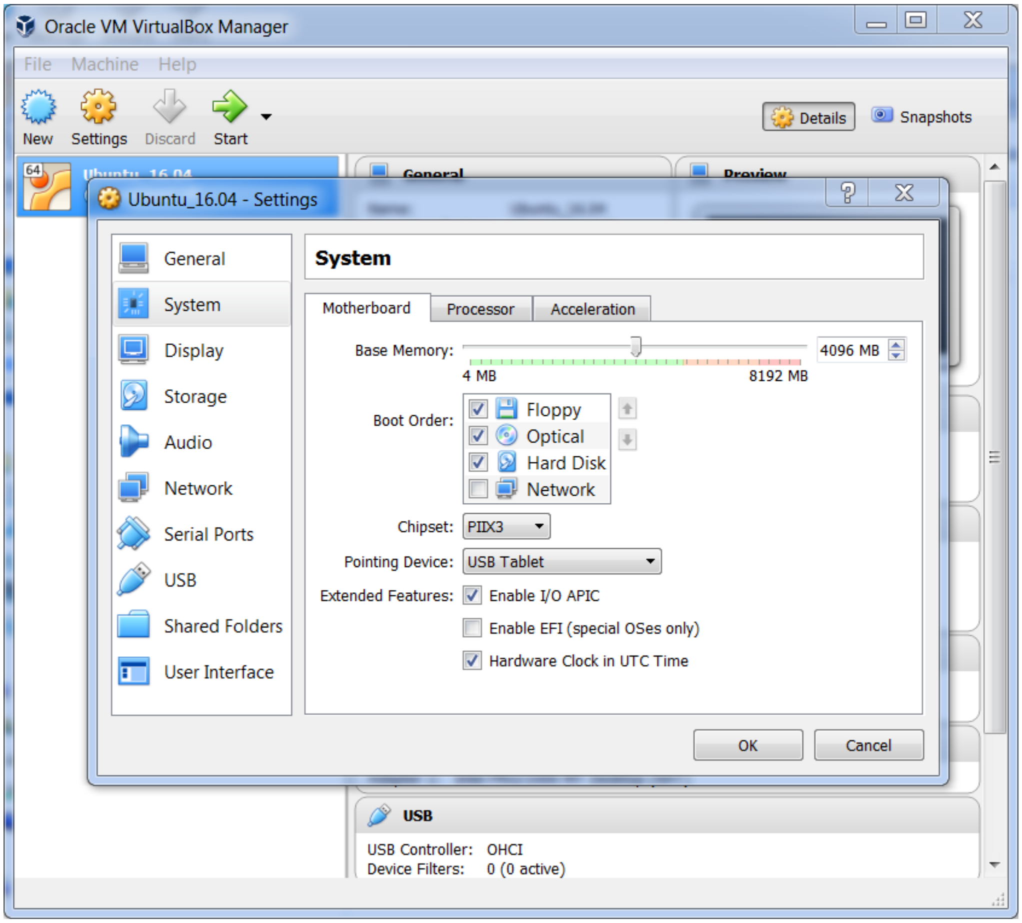
DVD ☑

USB Controller ☑

☐ Reinitialize the MAC address of all network cards

Appliance is not signed

By default, the VM will take up 4G of your physical memory. It is recommended that the VM memory usage does not exceed half of your total memory. To adjust memory usage, click “Settings” button, click “System” tab, and adjust “Base Memory”.



System

Base Memory: 4096 MB (4 MB – 8192 MB)

Boot Order: ☑ Floppy, ☑ Optical, ☑ Hard Disk, □ Network

Chipset: PIIX3

Pointing Device: USB Tablet

Extended Features: ☑ Enable I/O APIC

□ Enable EFI (special OSes only)

☑ Hardware Clock in UTC Time

USB

USB Controller: OHCI

Device Filters: 0 (0 active)

If there is any error while starting the VM related to USB, disable the USB controller from “Settings”, under the “USB” tab.

4. Now you can start the Ubuntu VM. By default, the username is “Spark” and password is “123456” (without quotation marks).

Appendix B. Install Spark-2.2.1 on Ubuntu-16.04 with JDK 8

1. Install JDK 8

Verify Java installation

```
$ java -version
```

If Java is not installed, we install it via the following commands.

```
$ sudo add-apt-repository ppa:webupd8team/java
$ sudo apt-get update && sudo apt-get install oracle-java8-installer
```

It may take some time to download the install. When it is done, set the path as follows.

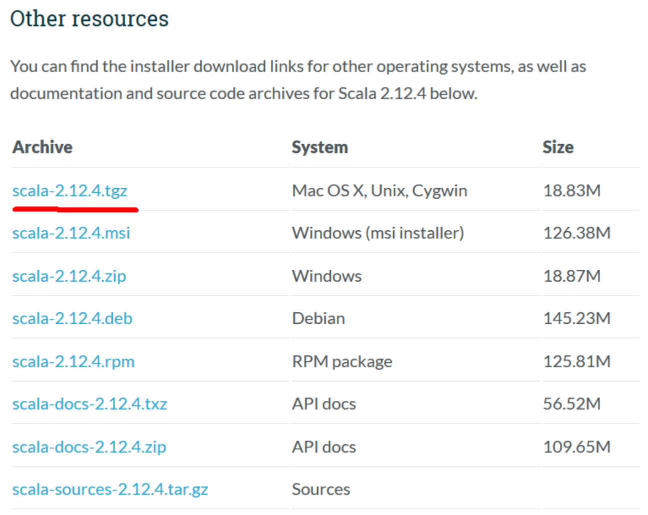
```
$ sudo gedit /etc/environment
```

Append the following line at the end of the file and save it.

JAVA_HOME="/usr/lib/jvm/java-8-oracle"

2. Install Scala

Download Scala in http://www.scala-lang.org/download/



Other resources

You can find the installer download links for other operating systems, as well as documentation and source code archives for Scala 2.12.4 below.

Archive	System	Size
scala-2.12.4.tgz	Mac OS X, Unix, Cygwin	18.83M
scala-2.12.4.msi	Windows (msi installer)	126.38M
scala-2.12.4.zip	Windows	18.87M
scala-2.12.4.deb	Debian	145.23M
scala-2.12.4.rpm	RPM package	125.81M
scala-docs-2.12.4.txz	API docs	56.52M
scala-docs-2.12.4.zip	API docs	109.65M
scala-sources-2.12.4.tar.gz	Sources	

```
$ cd /home/Spark/Downloads
$ tar xvf scala-2.12.4.tgz
```

Use the following commands to move the Scala files to the directory /usr/local/scala

```
$ su –
Password:
# cd /home/Spark/Downloads/
# mv scala-2.12.4 /usr/local/scala
# exit
```

If you have not set the password for root account, use the following command to set it.

```
$ sudo passwd
```

Set the path for Scala.

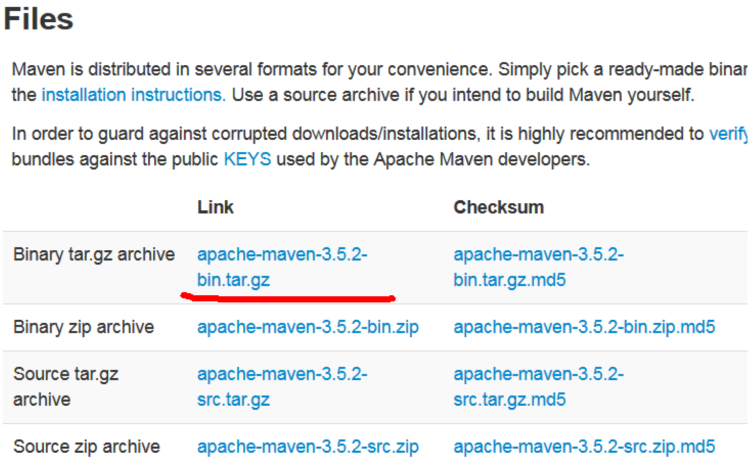
```
$ sudo gedit /etc/environment
```

Append the following clause to the end of PATH = “/usr/local/sbin:…..” in the file, and save it.

:/usr/local/scala/bin

3. Install Maven (to compile java files)

Download Maven from https://maven.apache.org/download.cgi



Files

Maven is distributed in several formats for your convenience. Simply pick a ready-made binar the installation instructions. Use a source archive if you intend to build Maven yourself.

In order to guard against corrupted downloads/installations, it is highly recommended to verify bundles against the public KEYS used by the Apache Maven developers.

	Link	Checksum
Binary tar.gz archive	apache-maven-3.5.2-bin.tar.gz	apache-maven-3.5.2-bin.tar.gz.md5
Binary zip archive	apache-maven-3.5.2-bin.zip	apache-maven-3.5.2-bin.zip.md5
Source tar.gz archive	apache-maven-3.5.2-src.tar.gz	apache-maven-3.5.2-src.tar.gz.md5
Source zip archive	apache-maven-3.5.2-src.zip	apache-maven-3.5.2-src.zip.md5

Extract Maven files.

```
$ cd /home/Spark/Downloads
$ tar xvf apache-maven-3.5.2-bin.tar.gz
```

Use the following commands to move the Maven files to the directory /usr/local/maven

```
$ su –
Password:
# cd /home/Spark/Downloads/
# mv apache-maven-3.5.2 /usr/local/maven
# exit
```

Set the path for Maven.

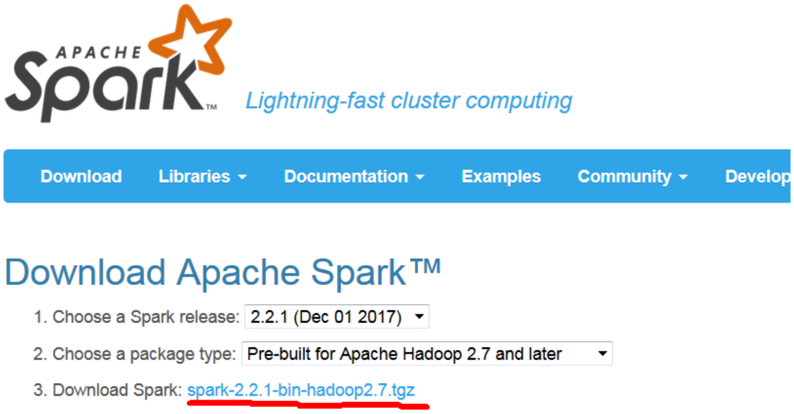
```
$ sudo gedit /etc/environment
```

Append the following clause at the end of PATH = “/usr/local/sbin:…..” in the file, and save it.

:/usr/local/maven/bin

4. Install Spark

Download Spark from https://spark.apache.org/downloads.html



Download Apache Spark™

1. Choose a Spark release: 2.2.1 (Dec 01 2017)
2. Choose a package type: Pre-built for Apache Hadoop 2.7 and later
3. Download Spark: spark-2.2.1-bin-hadoop2.7.tgz

Extract the file

```
$ cd /home/Spark/Downloads
$ tar xvf spark-2.2.1-bin-hadoop2.7.tgz
```

Move the Spark files to the directory /usr/local/spark

```
$ su –
Password:
# cd /home/Spark/Downloads/
# mv spark-2.2.1-bin-hadoop2.7 /usr/local/spark
# exit
```

Set the path for Spark.

```
$ sudo gedit /etc/environment
```

Append the following clause at the end of PATH = “/usr/local/sbin:…..”, then save it.

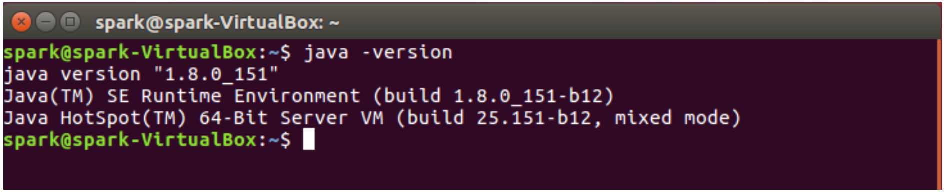
:/usr/local/spark/bin

Now, restart the system to make those changes work!

5. Verify the Software Installations

```
$ java -version
```

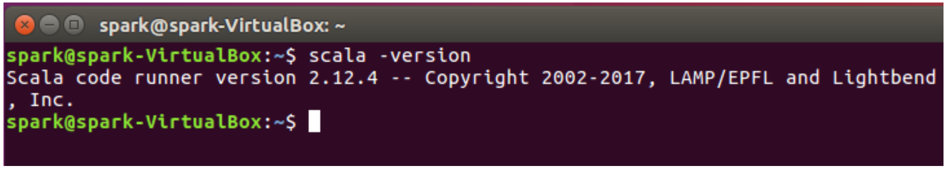
If Java is installed successfully then you will find the following output.



```
spark@spark-VirtualBox:~$ java -version
java version "1.8.0_151"
Java(TM) SE Runtime Environment (build 1.8.0_151-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.151-b12, mixed mode)
spark@spark-VirtualBox:~$
```

```
$ scala -version
```

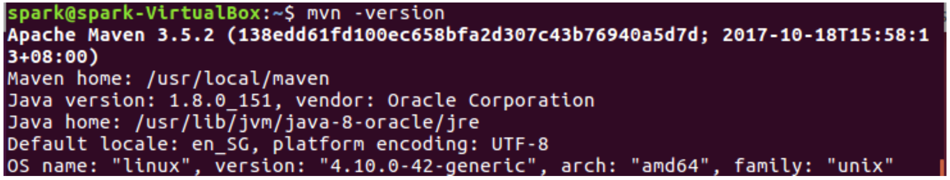
If Scala is installed successfully then you will find the following output.



```
spark@spark-VirtualBox:~$ scala -version
Scala code runner version 2.12.4 -- Copyright 2002-2017, LAMP/EPFL and Lightbend
, Inc.
spark@spark-VirtualBox:~$
```

```
$ mvn -version
```

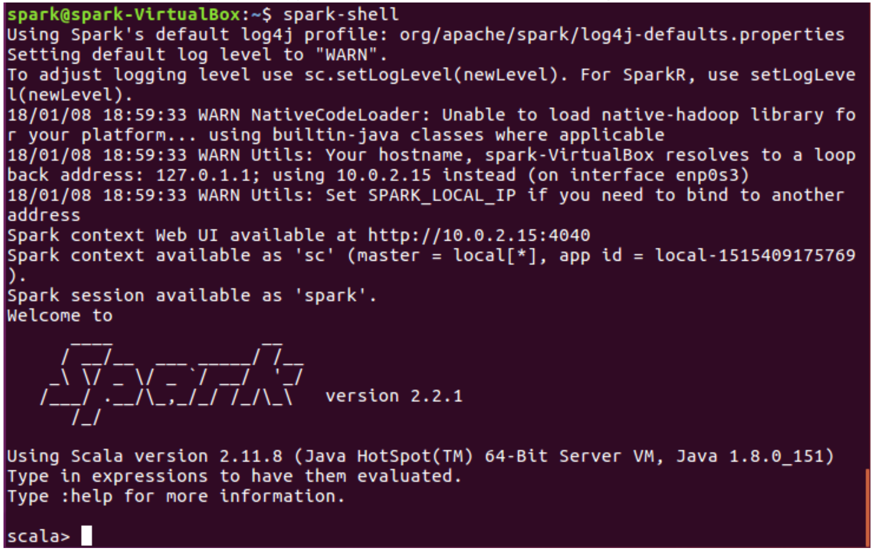
If Maven is installed successfully then you will find the following output.



```
spark@spark-VirtualBox:~$ mvn -version
Apache Maven 3.5.2 (138edd61fd100ec658bfa2d307c43b76940a5d7d; 2017-10-18T15:58:1
3+08:00)
Maven home: /usr/local/maven
Java version: 1.8.0_151, vendor: Oracle Corporation
Java home: /usr/lib/jvm/java-8-oracle/jre
Default locale: en_SG, platform encoding: UTF-8
OS name: "linux", version: "4.10.0-42-generic", arch: "amd64", family: "unix"
```

```
$ spark-shell
```

If spark is installed successfully then you will find the following output.



```
spark@spark-VirtualBox:~$ spark-shell
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
18/01/08 18:59:33 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
18/01/08 18:59:33 WARN Utils: Your hostname, spark-VirtualBox resolves to a loop
back address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
18/01/08 18:59:33 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1515409175769
).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.2.1
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_151)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

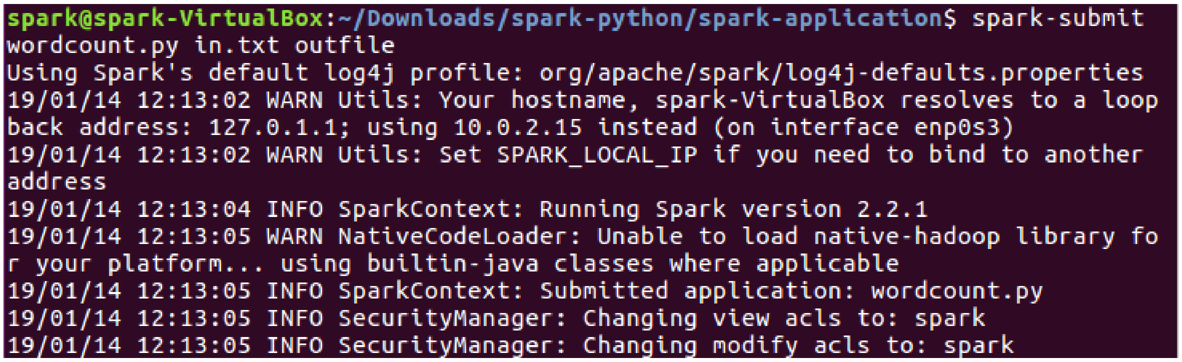
Appendix C. My First Spark Program (with Python)

1. Download the example files from Lab1 folder (in.txt, wordcount.py).

2. Create a new folder named ”spark-application” with the files you downloaded. (in.txt, wordcount.py).

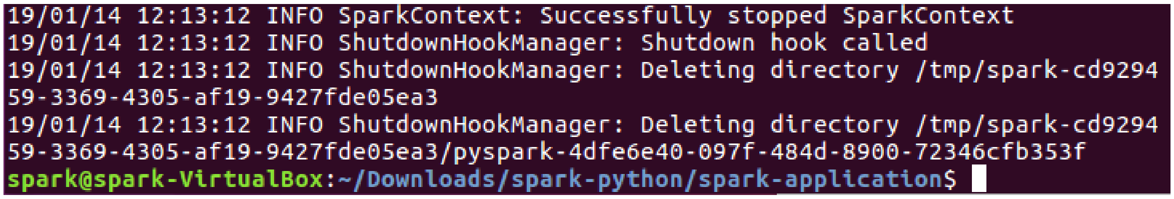
3. To execute the Spark program, using the following command under the folder “…/spark-application/”.

```
$ spark-submit wordcount.py in.txt outfile
```



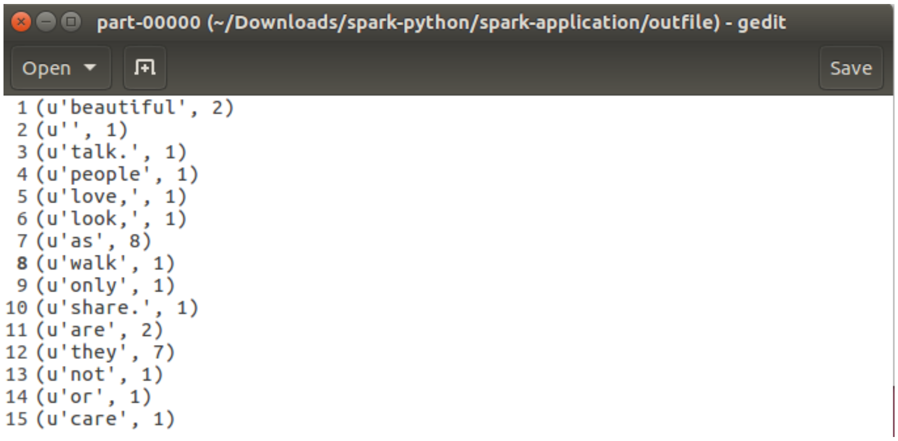
```
spark@spark-VirtualBox:~/Downloads/spark-python/spark-application$ spark-submit
wordcount.py in.txt outfile
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
19/01/14 12:13:02 WARN Utils: Your hostname, spark-VirtualBox resolves to a loop
back address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
19/01/14 12:13:02 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
19/01/14 12:13:04 INFO SparkContext: Running Spark version 2.2.1
19/01/14 12:13:05 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
19/01/14 12:13:05 INFO SparkContext: Submitted application: wordcount.py
19/01/14 12:13:05 INFO SecurityManager: Changing view acls to: spark
19/01/14 12:13:05 INFO SecurityManager: Changing modify acls to: spark
```

...



```
19/01/14 12:13:12 INFO SparkContext: Successfully stopped SparkContext
19/01/14 12:13:12 INFO ShutdownHookManager: Shutdown hook called
19/01/14 12:13:12 INFO ShutdownHookManager: Deleting directory /tmp/spark-cd9294
59-3369-4305-af19-9427fde05ea3
19/01/14 12:13:12 INFO ShutdownHookManager: Deleting directory /tmp/spark-cd9294
59-3369-4305-af19-9427fde05ea3/pyspark-4dfe6e40-097f-484d-8900-72346cfb353f
spark@spark-VirtualBox:~/Downloads/spark-python/spark-application$
```

You can see a folder named *outfile* generated under ”.” directory. The result is in the inside file named *part-00000*.


part-00000 (~/Downloads/spark-python/spark-application/outfile) - gedit
Open
Save

```
(u'beautiful', 2)
(u'', 1)
(u'talk.', 1)
(u'people', 1)
(u'love,', 1)
(u'look,', 1)
(u'as', 8)
(u'walk', 1)
(u'only', 1)
(u'share.', 1)
(u'are', 2)
(u'they', 7)
(u'not', 1)
(u'or', 1)
(u'care', 1)
```