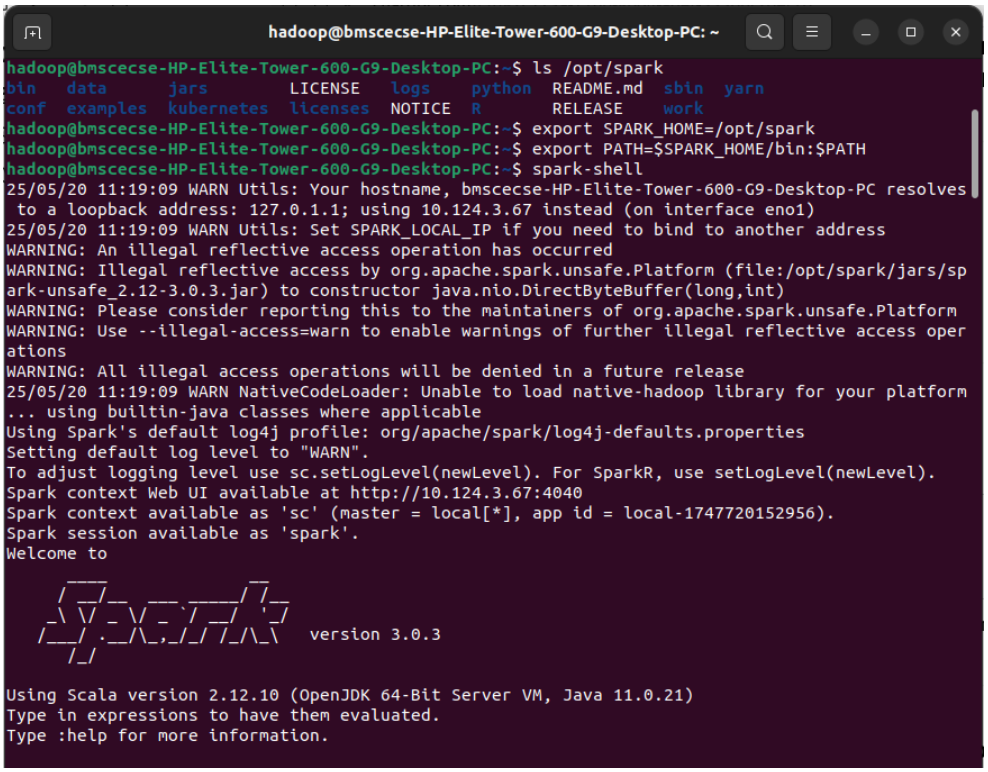
## LABORATORY PROGRAM – 8

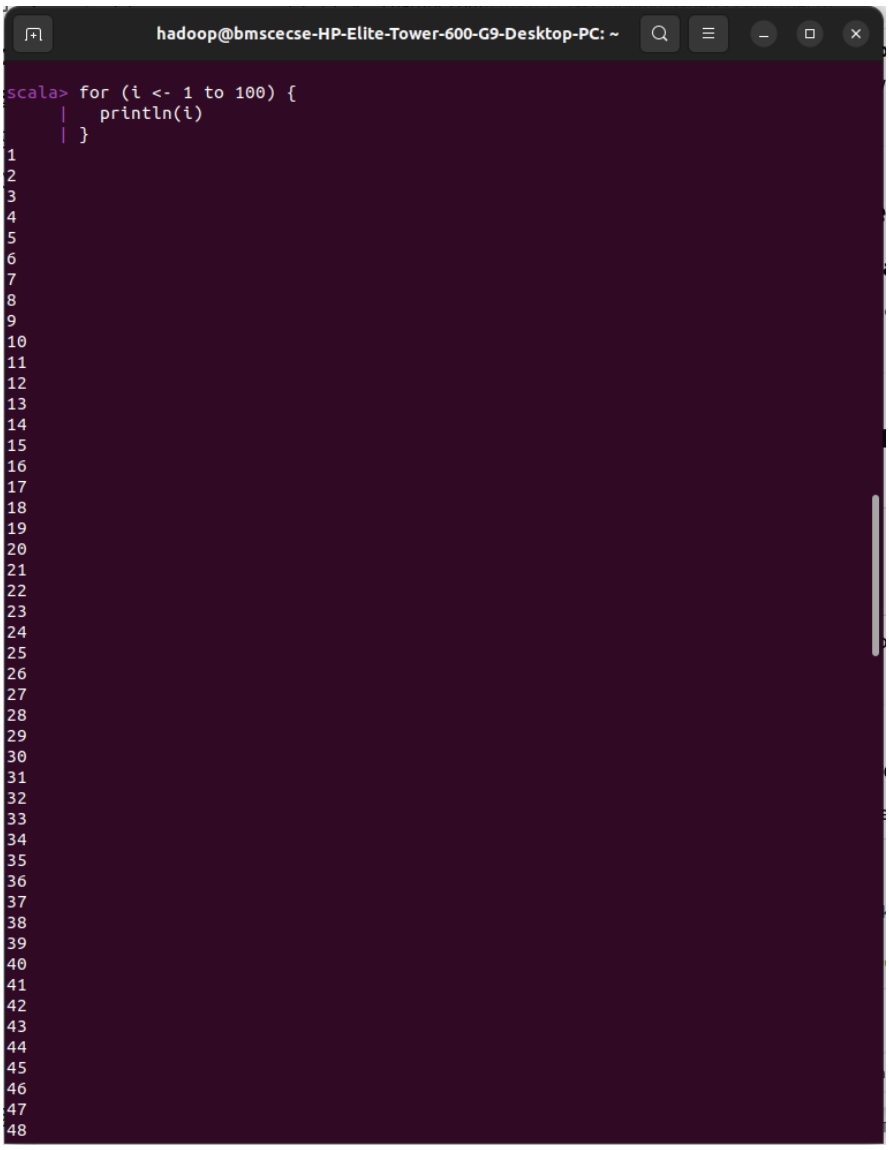
## Scala Program to print numbers from 1 to 100

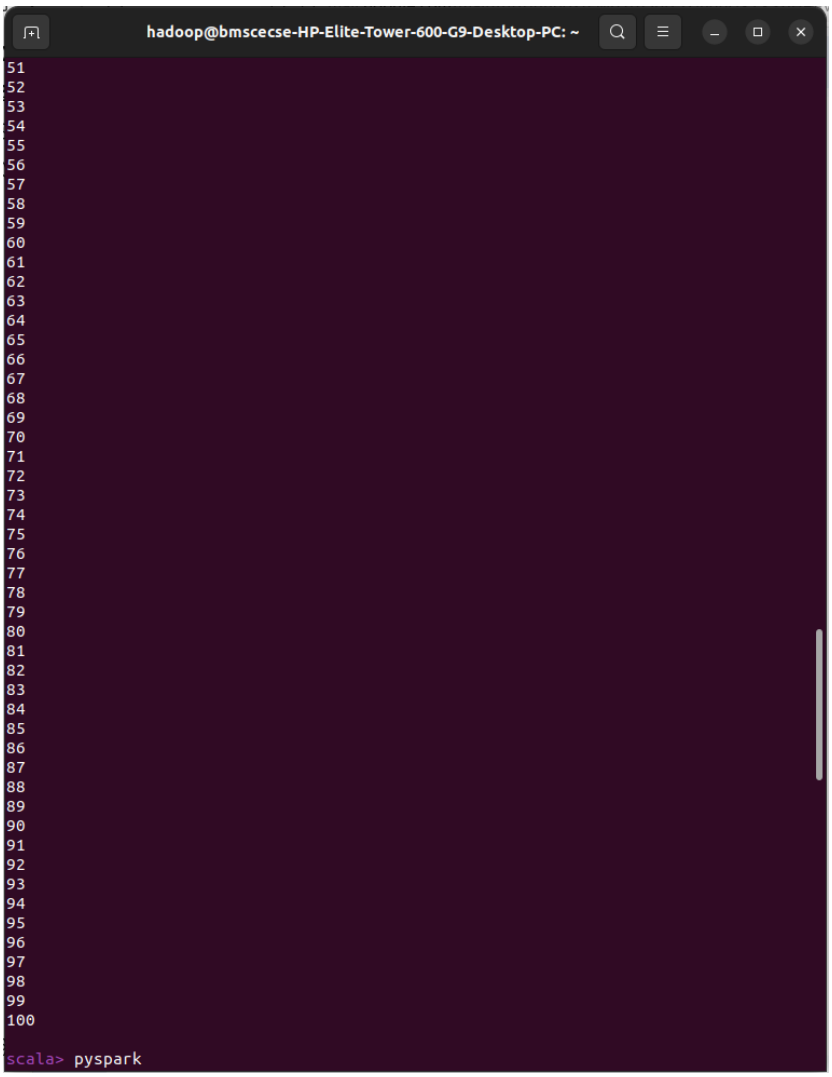
```
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
import re

# Initialize SparkContext and StreamingContext (batch interval = 5 sec)
sc = SparkContext("local[2]", "TextCleanStreamingApp")
ssc = StreamingContext(sc, 5)

# Set up stop words and lemmatizer
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()

# Connect to the socket stream
lines = ssc.socketTextStream("localhost", 9999)

# Text cleaning function
def clean_text(line):
    # Remove non-alphabetic characters and lower the case
    line = re.sub(r'[^a-zA-Z\s]', '', line)
    line = line.lower()
    # Tokenize
    tokens = word_tokenize(line)
    # Remove stop words and lemmatize
    cleaned_tokens = [lemmatizer.lemmatize(word) for word in tokens if word not in stop_words]
```

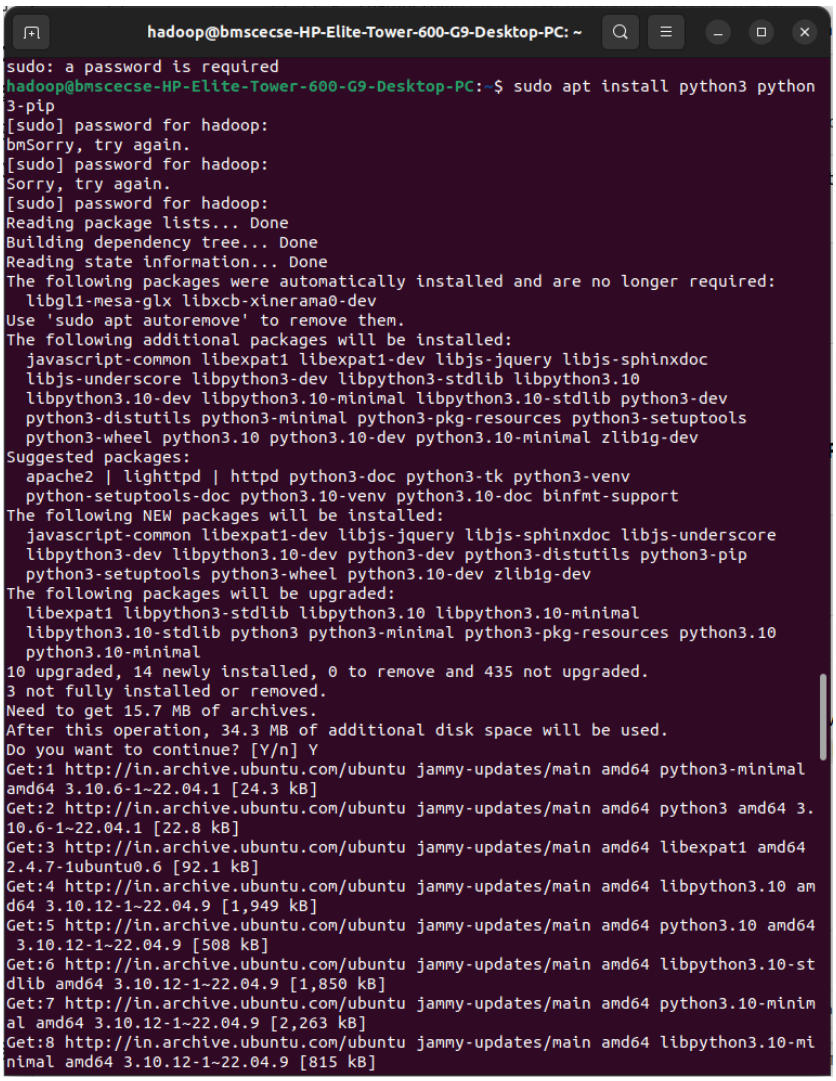
```
    return ' '.join(cleaned_tokens)

# Apply cleaning to each line
cleaned_lines = lines.map(clean_text)

# Print the cleaned lines
cleaned_lines.pprint()

# Start streaming
ssc.start()
ssc.awaitTermination()
```

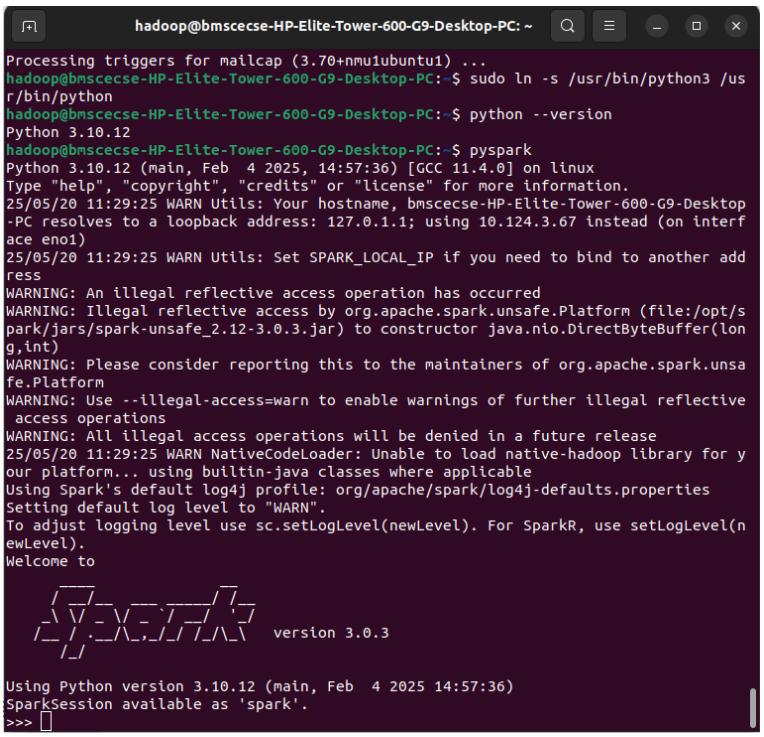
**OBSERVATION**

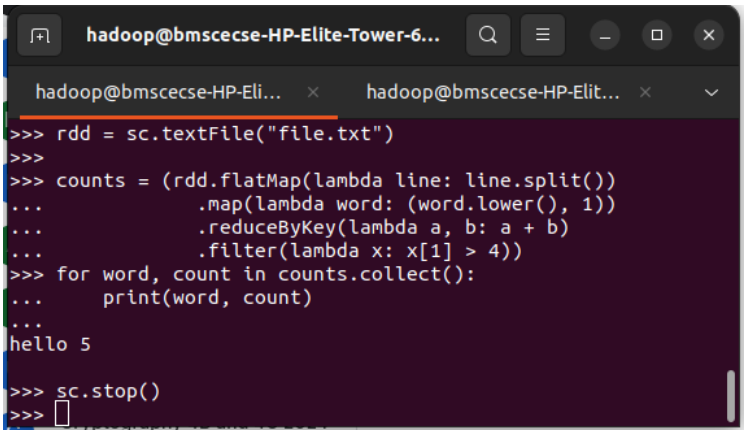



hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~
scala> for (i <- 1 to 100) {
| println(i)
| }
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48


hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
scala> pyspark


hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~
sudo: a password is required
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ sudo apt install python3 python
3-pip
[sudo] password for hadoop:
bmSorry, try again.
[sudo] password for hadoop:
Sorry, try again.
[sudo] password for hadoop:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following packages were automatically installed and are no longer required:
libgl1-mesa-glx libxcb-xinerama0-dev
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
javascript-common libexpat1 libexpat1-dev libjs-jquery libjs-sphinxdoc
libjs-underscore libpython3-dev libpython3-stdlib libpython3.10
libpython3.10-dev libpython3.10-minimal libpython3.10-stdlib python3-dev
python3-distutils python3-minimal python3-pkg-resources python3-setuptools
python3-wheel python3.10 python3.10-dev python3.10-minimal zlib1g-dev
Suggested packages:
apache2 | lighttpd | httpd python3-doc python3-tk python3-venv
python-setuptools-doc python3.10-venv python3.10-doc binfmt-support
The following NEW packages will be installed:
javascript-common libexpat1-dev libjs-jquery libjs-sphinxdoc libjs-underscore
libpython3-dev libpython3.10-dev python3-dev python3-distutils python3-pip
python3-setuptools python3-wheel python3.10-dev zlib1g-dev
The following packages will be upgraded:
libexpat1 libpython3-stdlib libpython3.10 libpython3.10-minimal
libpython3.10-stdlib python3 python3-minimal python3-pkg-resources python3.10
python3.10-minimal
10 upgraded, 14 newly installed, 0 to remove and 435 not upgraded.
3 not fully installed or removed.
Need to get 15.7 MB of archives.
After this operation, 34.3 MB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://in.archive.ubuntu.com/ubuntu jammy-updates/main amd64 python3-minimal
amd64 3.10.6-1~22.04.1 [24.3 kB]
Get:2 http://in.archive.ubuntu.com/ubuntu jammy-updates/main amd64 python3 amd64 3.
10.6-1~22.04.1 [22.8 kB]
Get:3 http://in.archive.ubuntu.com/ubuntu jammy-updates/main amd64 libexpat1 amd64
2.4.7-1ubuntu0.6 [92.1 kB]
Get:4 http://in.archive.ubuntu.com/ubuntu jammy-updates/main amd64 libpython3.10 am
d64 3.10.12-1~22.04.9 [1,949 kB]
Get:5 http://in.archive.ubuntu.com/ubuntu jammy-updates/main amd64 python3.10 amd64
3.10.12-1~22.04.9 [508 kB]
Get:6 http://in.archive.ubuntu.com/ubuntu jammy-updates/main amd64 libpython3.10-st
dlib amd64 3.10.12-1~22.04.9 [1,850 kB]
Get:7 http://in.archive.ubuntu.com/ubuntu jammy-updates/main amd64 python3.10-minim
al amd64 3.10.12-1~22.04.9 [2,263 kB]
Get:8 http://in.archive.ubuntu.com/ubuntu jammy-updates/main amd64 libpython3.10-mi
nimal amd64 3.10.12-1~22.04.9 [815 kB]


hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~
Processing triggers for mailcap (3.70+nmu1ubuntu1) ...
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ sudo ln -s /usr/bin/python3 /usr/bin/python
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ python --version
Python 3.10.12
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ pyspark
Python 3.10.12 (main, Feb 4 2025, 14:57:36) [GCC 11.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
25/05/20 11:29:25 WARN Utils: Your hostname, bmscecse-HP-Elite-Tower-600-G9-Desktop-PC resolves to a loopback address: 127.0.1.1; using 10.124.3.67 instead (on interface eno1)
25/05/20 11:29:25 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/jars/spark-unsafe_2.12-3.0.3.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
25/05/20 11:29:25 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
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 setLogLevel(newLevel).
Welcome to
version 3.0.3
Using Python version 3.10.12 (main, Feb 4 2025 14:57:36)
SparkSession available as 'spark'.
>>>


hadoop@bmscecse-HP-Elite-Tower-6...
hadoop@bmscecse-HP-Eli...
hadoop@bmscecse-HP-Elit...
>>> rdd = sc.textFile("file.txt")
>>>
>>> counts = (rdd.flatMap(lambda line: line.split())
... .map(lambda word: (word.lower(), 1))
... .reduceByKey(lambda a, b: a + b)
... .filter(lambda x: x[1] > 4))
>>> for word, count in counts.collect():
... print(word, count)
...
hello 5
>>> sc.stop()
>>>