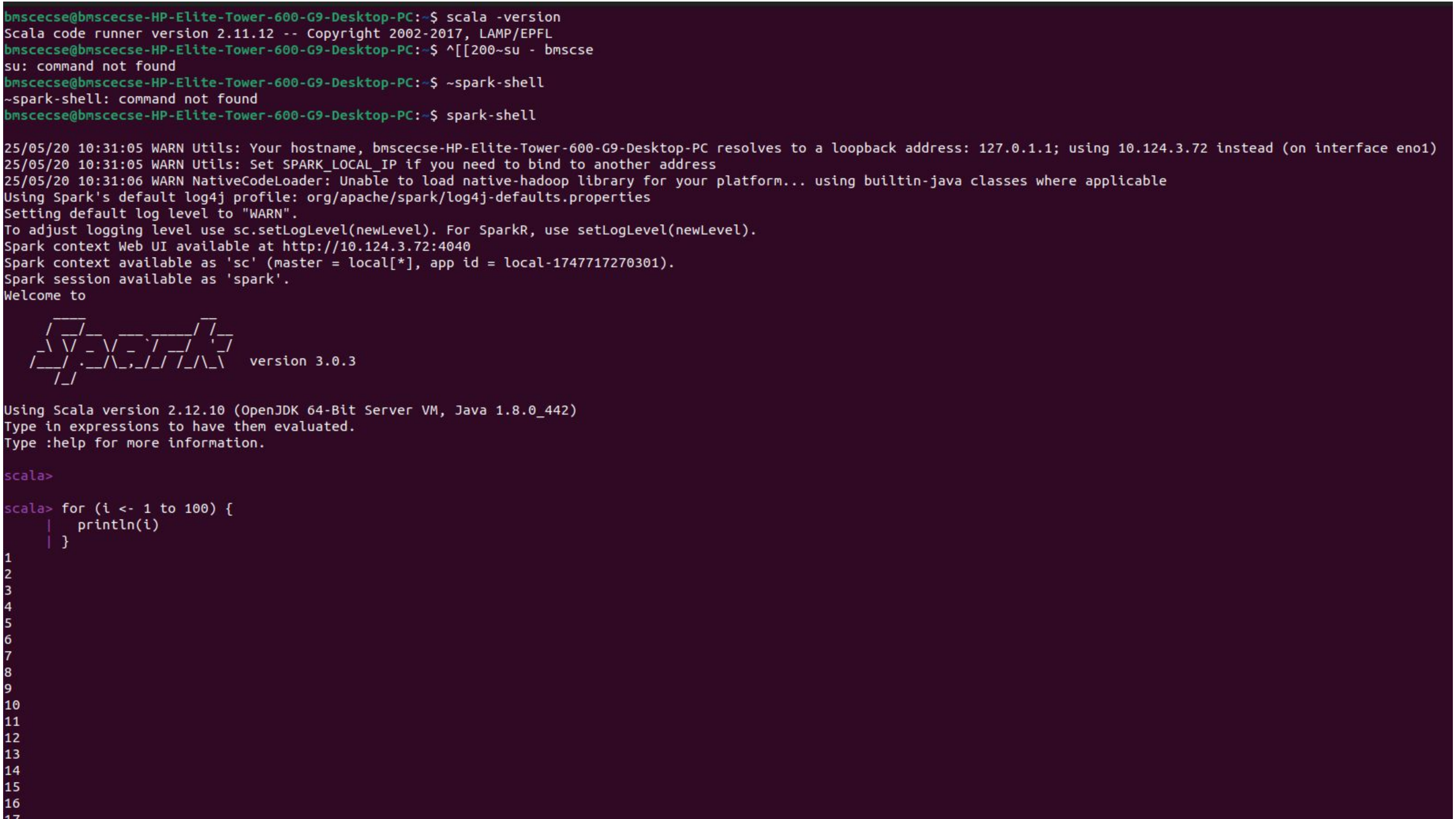
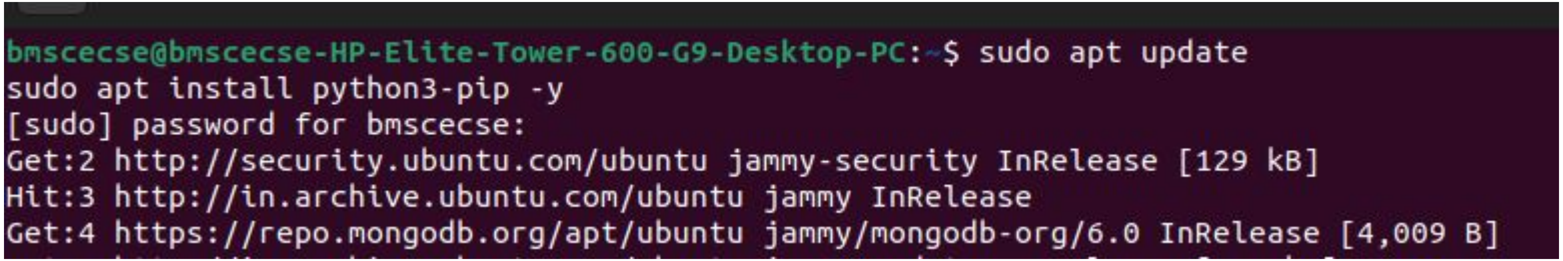
Scala

1BM22CS158

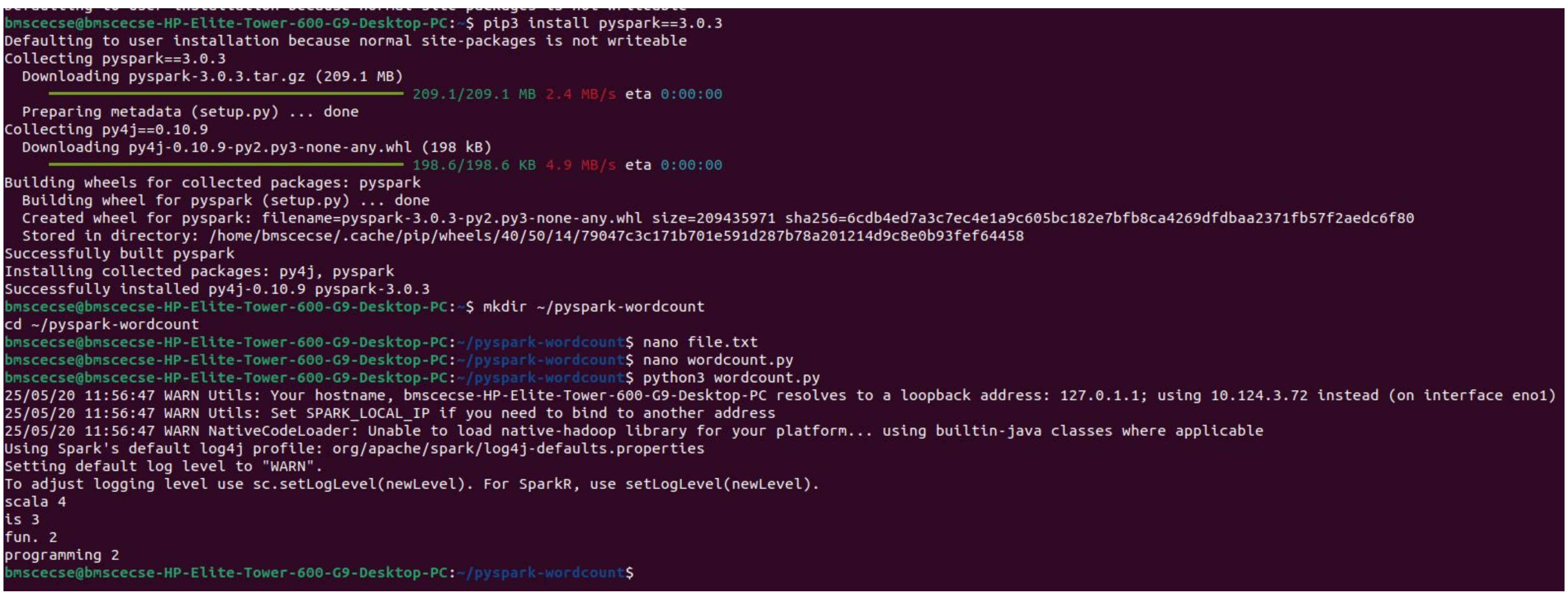


```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ scala -version
Scala code runner version 2.11.12 -- Copyright 2002-2017, LAMP/EPFL
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ ^[[200~su - bmscse
su: command not found
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ ~spark-shell
~spark-shell: command not found
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ spark-shell

25/05/20 10:31:05 WARN Utils: Your hostname, bmscecse-HP-Elite-Tower-600-G9-Desktop-PC resolves to a loopback address: 127.0.1.1; using 10.124.3.72 instead (on interface eno1)
25/05/20 10:31:05 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
25/05/20 10:31:06 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).
Spark context Web UI available at http://10.124.3.72:4040
Spark context available as 'sc' (master = local[*], app id = local-1747717270301).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.0.3
      /_/

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 1.8.0_442)
Type in expressions to have them evaluated.
Type :help for more information.

scala>

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
```

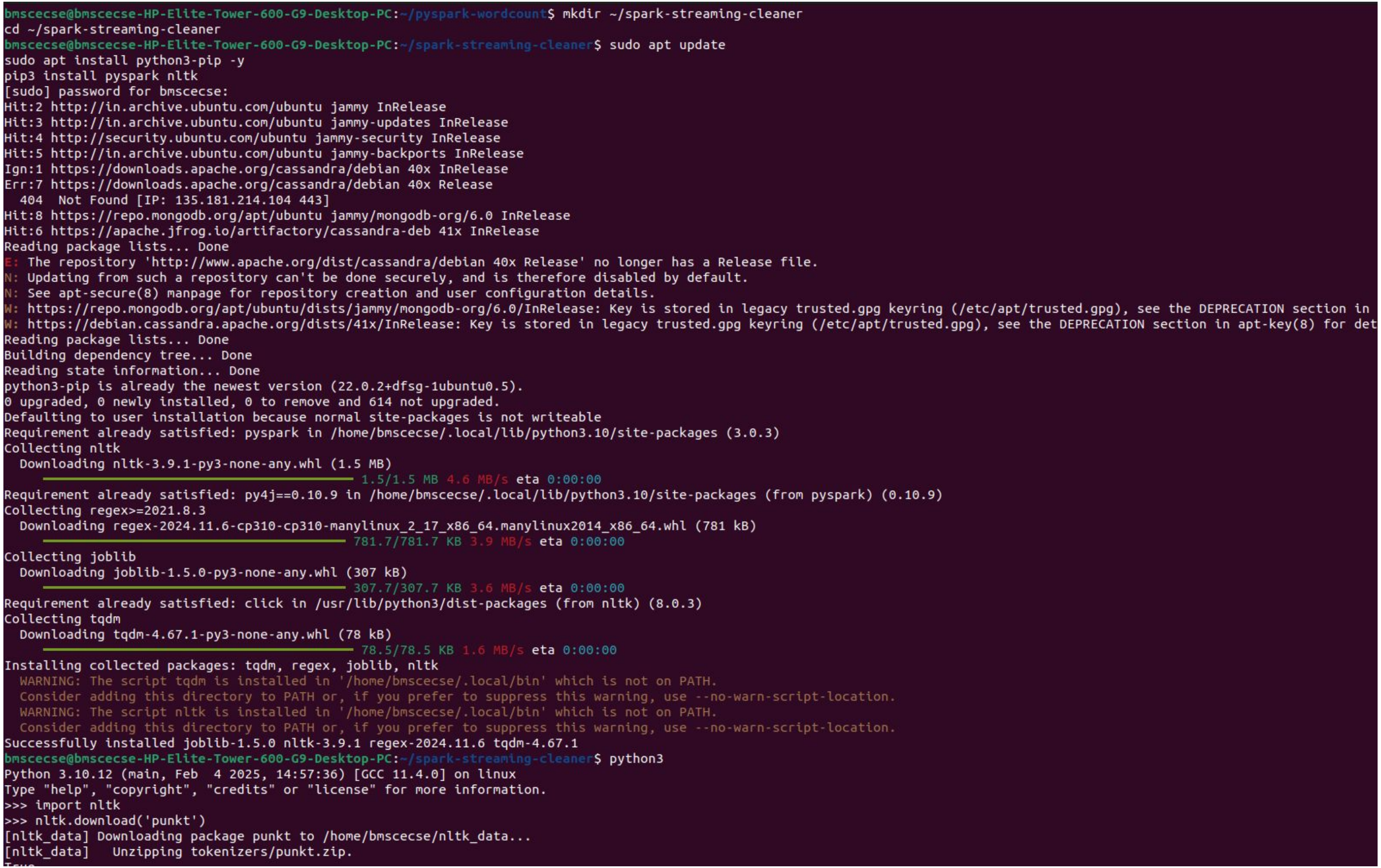
```
scala> :quit
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$
```



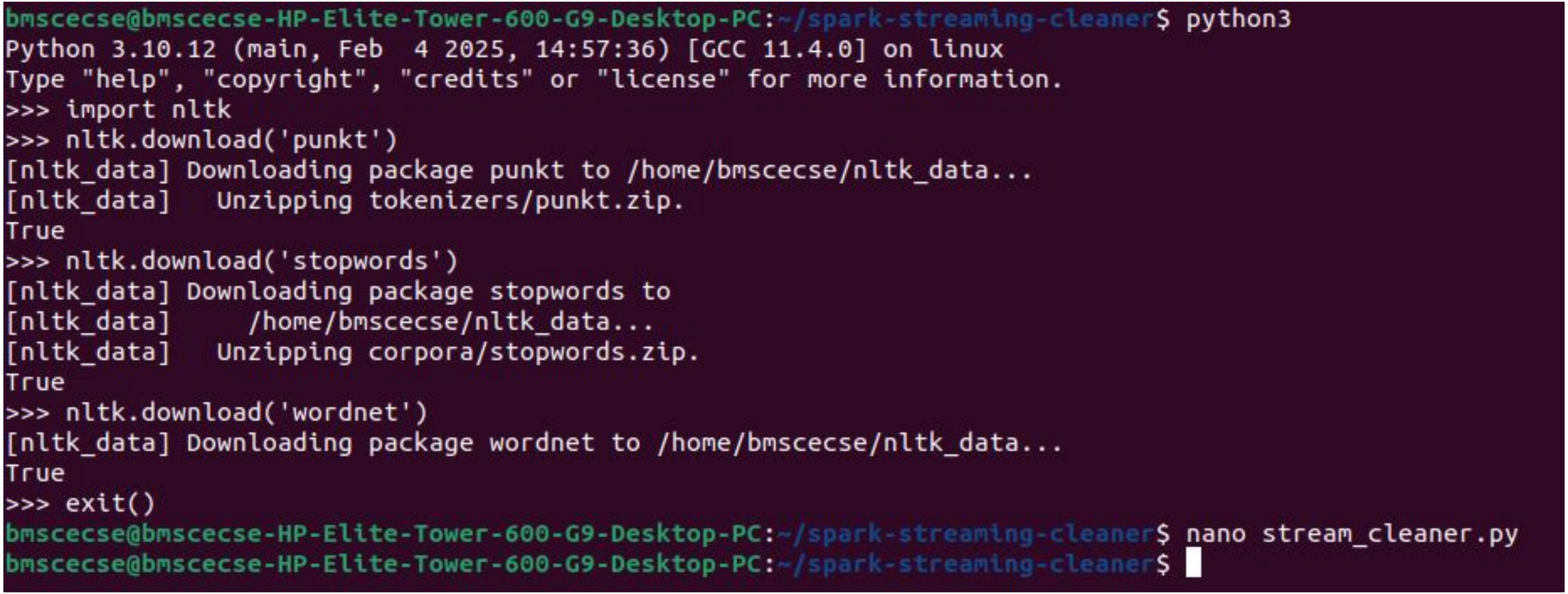
```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ sudo apt update
sudo apt install python3-pip -y
[sudo] password for bmscecse:
Get:2 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Hit:3 http://in.archive.ubuntu.com/ubuntu jammy InRelease
Get:4 https://repo.mongodb.org/apt/ubuntu jammy/mongodb-org/6.0 InRelease [4,009 B]
```



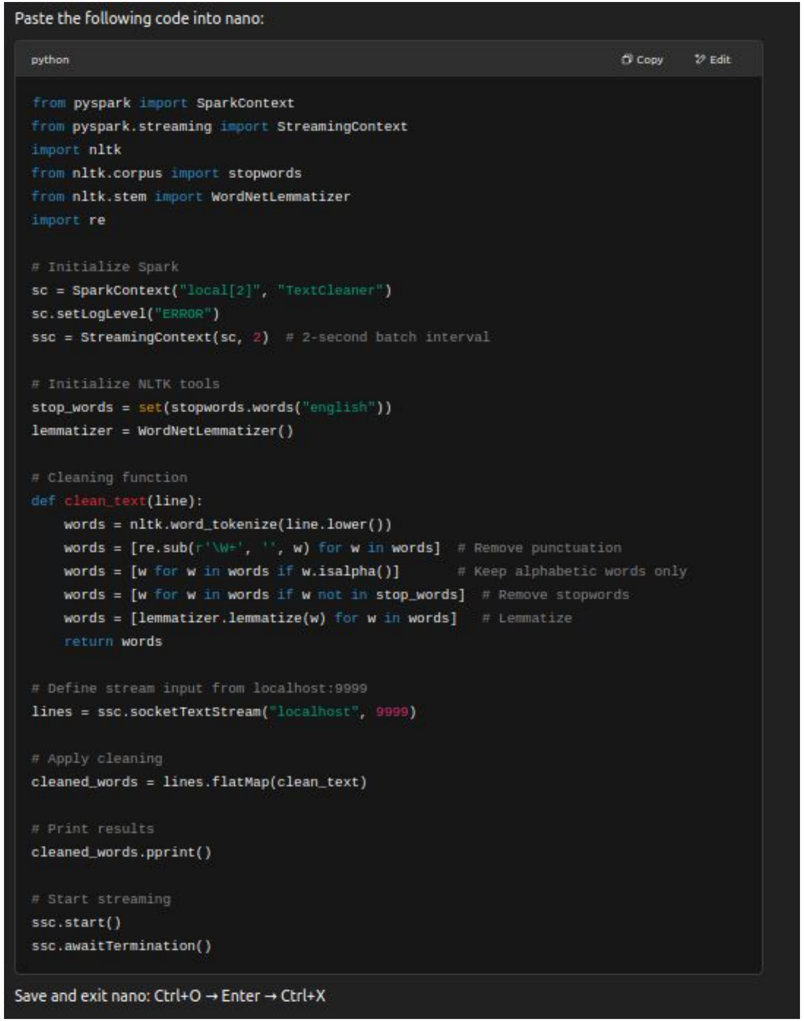
```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ pip3 install pyspark==3.0.3
Defaulting to user installation because normal site-packages is not writeable
Collecting pyspark==3.0.3
  Downloading pyspark-3.0.3.tar.gz (209.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 209.1/209.1 MB 2.4 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting py4j==0.10.9
  Downloading py4j-0.10.9-py2.py3-none-any.whl (198 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 198.6/198.6 KB 4.9 MB/s eta 0:00:00
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.0.3-py2.py3-none-any.whl size=209435971 sha256=6cdb4ed7a3c7ec4e1a9c605bc182e7bfb8ca4269dfdbaa2371fb57f2aedc6f80
  Stored in directory: /home/bmscecse/.cache/pip/wheels/40/50/14/79047c3c171b701e591d287b78a201214d9c8e0b93fef64458
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9 pyspark-3.0.3
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ mkdir ~/pyspark-wordcount
cd ~/pyspark-wordcount
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/pyspark-wordcount$ nano file.txt
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/pyspark-wordcount$ nano wordcount.py
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/pyspark-wordcount$ python3 wordcount.py
25/05/20 11:56:47 WARN Utils: Your hostname, bmscecse-HP-Elite-Tower-600-G9-Desktop-PC resolves to a loopback address: 127.0.1.1; using 10.124.3.72 instead (on interface eno1)
25/05/20 11:56:47 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
25/05/20 11:56:47 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).
scala 4
is 3
fun. 2
programming 2
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/pyspark-wordcount$
```



```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/pyspark-wordcount$ mkdir ~/spark-streaming-cleaner
cd ~/spark-streaming-cleaner
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ sudo apt update
sudo apt install python3-pip -y
pip3 install pyspark nltk
[sudo] password for bmscecse:
Hit:2 http://in.archive.ubuntu.com/ubuntu jammy InRelease
Hit:3 http://in.archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:4 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:5 http://in.archive.ubuntu.com/ubuntu jammy-backports InRelease
Ign:1 https://downloads.apache.org/cassandra/debian 40x InRelease
Err:7 https://downloads.apache.org/cassandra/debian 40x Release
  404  Not Found [IP: 135.181.214.104 443]
Hit:8 https://repo.mongodb.org/apt/ubuntu jammy/mongodb-org/6.0 InRelease
Hit:6 https://apache.jfrog.io/artifactory/cassandra-deb 41x InRelease
Reading package lists... Done
E: The repository 'http://www.apache.org/dist/cassandra/debian 40x Release' no longer has a Release file.
N: Updating from such a repository can't be done securely, and is therefore disabled by default.
N: See apt-secure(8) manpage for repository creation and user configuration details.
W: https://repo.mongodb.org/apt/ubuntu/dists/jammy/mongodb-org/6.0/InRelease: Key is stored in legacy trusted.gpg keyring (/etc/apt/trusted.gpg), see the DEPRECATION section in
W: https://debian.cassandra.apache.org/dists/41x/InRelease: Key is stored in legacy trusted.gpg keyring (/etc/apt/trusted.gpg), see the DEPRECATION section in apt-key(8) for det
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
python3-pip is already the newest version (22.0.2+dfsg-1ubuntu0.5).
0 upgraded, 0 newly installed, 0 to remove and 614 not upgraded.
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pyspark in /home/bmscecse/.local/lib/python3.10/site-packages (3.0.3)
Collecting nltk
  Downloading nltk-3.9.1-py3-none-any.whl (1.5 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.5/1.5 MB 4.6 MB/s eta 0:00:00
Requirement already satisfied: py4j==0.10.9 in /home/bmscecse/.local/lib/python3.10/site-packages (from pyspark) (0.10.9)
Collecting regex>=2021.8.3
  Downloading regex-2024.11.6-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (781 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 781.7/781.7 KB 3.9 MB/s eta 0:00:00
Collecting joblib
  Downloading joblib-1.5.0-py3-none-any.whl (307 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 307.7/307.7 KB 3.6 MB/s eta 0:00:00
Requirement already satisfied: click in /usr/lib/python3/dist-packages (from nltk) (8.0.3)
Collecting tqdm
  Downloading tqdm-4.67.1-py3-none-any.whl (78 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 78.5/78.5 KB 1.6 MB/s eta 0:00:00
Installing collected packages: tqdm, regex, joblib, nltk
  WARNING: The script tqdm is installed in '/home/bmscecse/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
  WARNING: The script nltk is installed in '/home/bmscecse/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed joblib-1.5.0 nltk-3.9.1 regex-2024.11.6 tqdm-4.67.1
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ python3
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.
>>> import nltk
>>> nltk.download('punkt')
[nltk_data] Downloading package punkt to /home/bmscecse/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```



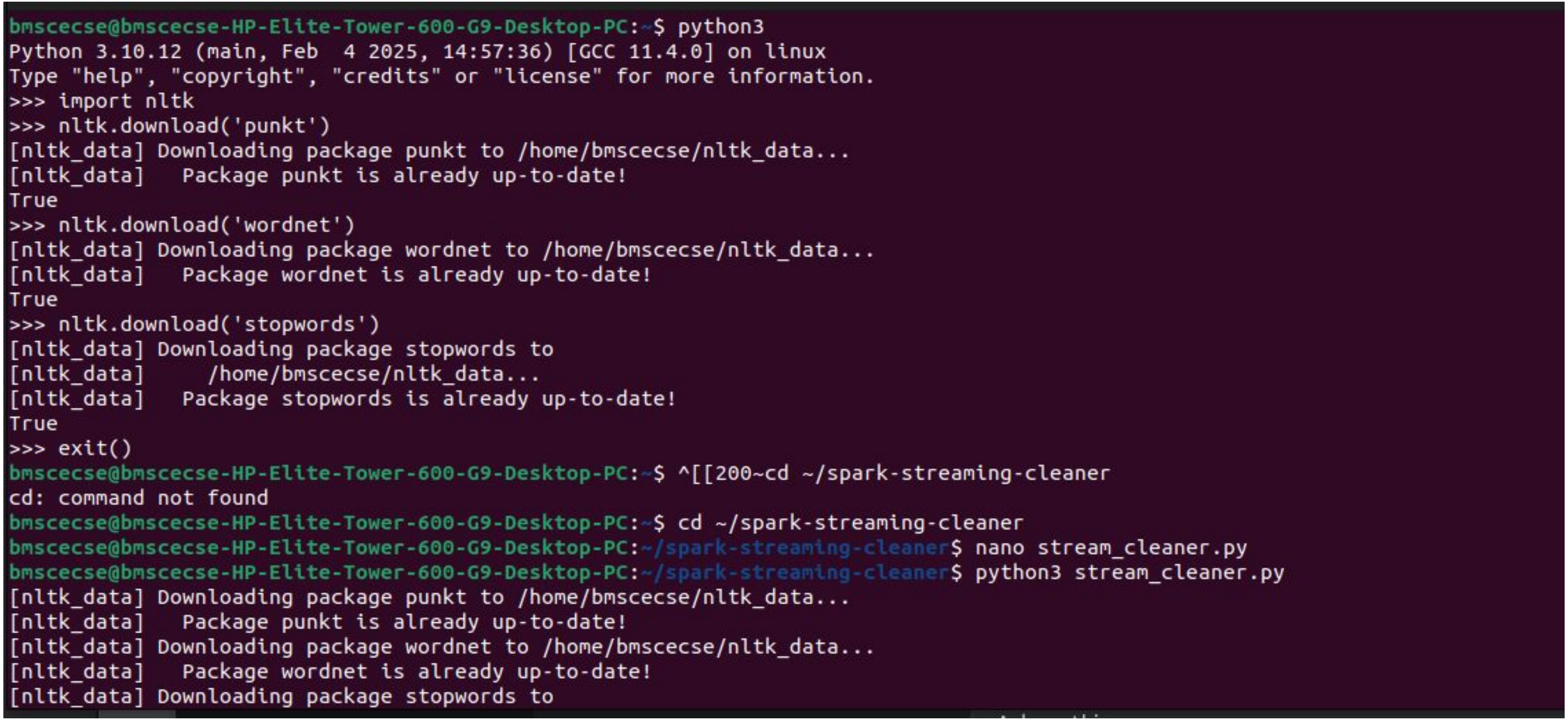
```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ python3
Python 3.10.12 (main, Feb  4 2025, 14:57:36) [GCC 11.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> nltk.download('punkt')
[nltk_data] Downloading package punkt to /home/bmscecse/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
>>> nltk.download('stopwords')
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/bmscecse/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
>>> nltk.download('wordnet')
[nltk_data] Downloading package wordnet to /home/bmscecse/nltk_data...
True
>>> exit()
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ nano stream_cleaner.py
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ 
```



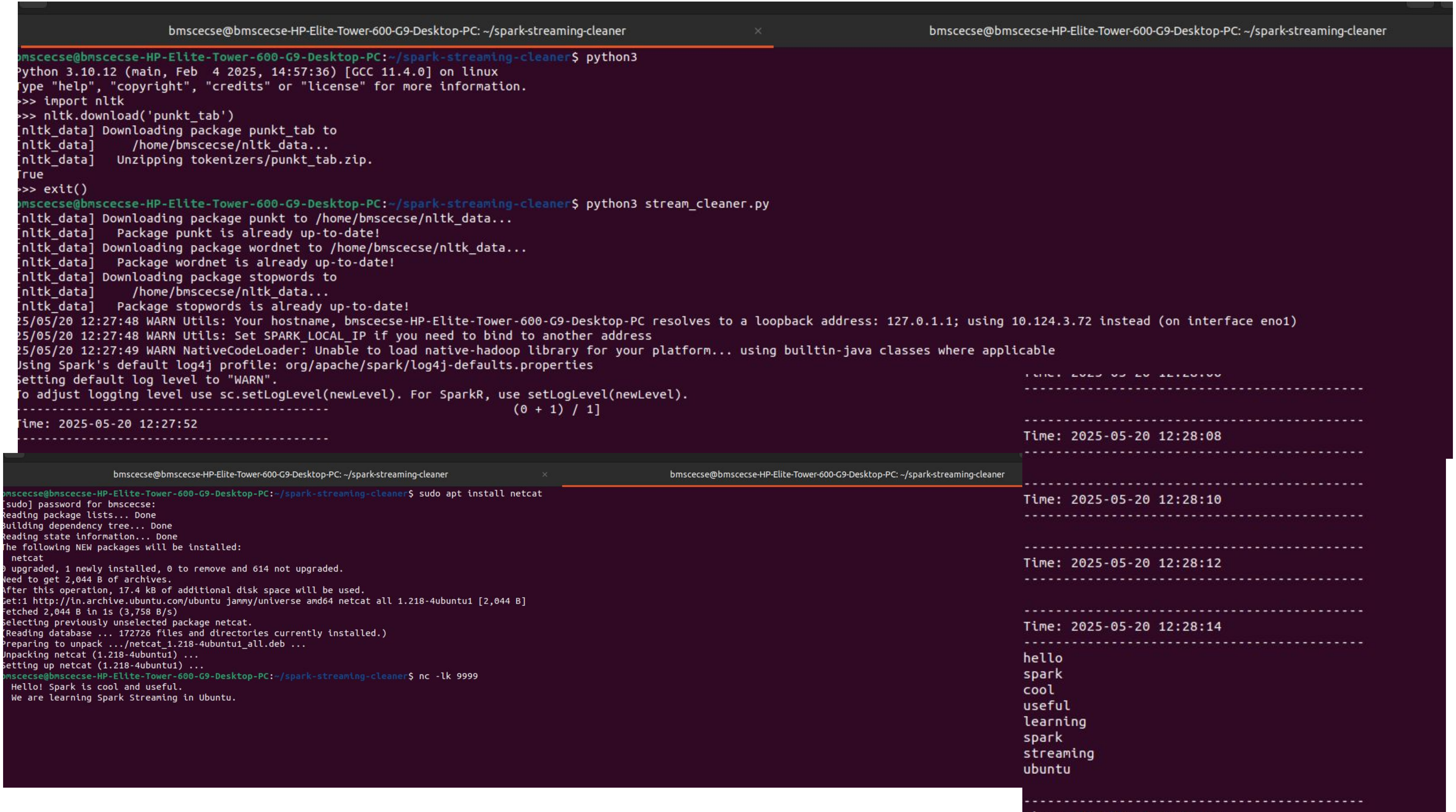
Paste the following code into nano:

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

# Initialize Spark
sc = SparkContext("local[2]", "TextCleaner")
sc.setLogLevel("ERROR")
ssc = StreamingContext(sc, 2)  # 2-second batch interval

# Initialize NLTK tools
stop_words = set(stopwords.words("english"))
lemmatizer = WordNetLemmatizer()

# Cleaning function
def clean_text(line):
    words = nltk.word_tokenize(line.lower())
    words = [re.sub(r'\W+', '', w) for w in words]  # Remove punctuation
    words = [w for w in words if w.isalpha()]       # Keep alphabetic words only
    words = [w for w in words if w not in stop_words]  # Remove stopwords
    words = [lemmatizer.lemmatize(w) for w in words]   # Lemmatize
    return words

# Define stream input from localhost:9999
lines = ssc.socketTextStream("localhost", 9999)

# Apply cleaning
cleaned_words = lines.flatMap(clean_text)

# Print results
cleaned_words.pprint()

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

Save and exit nano: Ctrl+O → Enter → Ctrl+X



```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ python3
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.
>>> import nltk
>>> nltk.download('punkt')
[nltk_data] Downloading package punkt to /home/bmscecse/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
True
>>> nltk.download('wordnet')
[nltk_data] Downloading package wordnet to /home/bmscecse/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
True
>>> nltk.download('stopwords')
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/bmscecse/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True
>>> exit()
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ ^[[200~cd ~/spark-streaming-cleaner
cd: command not found
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cd ~/spark-streaming-cleaner
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ nano stream_cleaner.py
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ python3 stream_cleaner.py
[nltk_data] Downloading package punkt to /home/bmscecse/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /home/bmscecse/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package stopwords to
```



bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~/spark-streaming-cleaner
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~/spark-streaming-cleaner

```
mscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ python3
ython 3.10.12 (main, Feb  4 2025, 14:57:36) [GCC 11.4.0] on linux
ype "help", "copyright", "credits" or "license" for more information.
>> import nltk
>> nltk.download('punkt_tab')
nltk_data] Downloading package punkt_tab to
nltk_data]     /home/bmscecse/nltk_data...
nltk_data]   Unzipping tokenizers/punkt_tab.zip.
rue
>> exit()
mscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ python3 stream_cleaner.py
nltk_data] Downloading package punkt to /home/bmscecse/nltk_data...
nltk_data]   Package punkt is already up-to-date!
nltk_data] Downloading package wordnet to /home/bmscecse/nltk_data...
nltk_data]   Package wordnet is already up-to-date!
nltk_data] Downloading package stopwords to
nltk_data]     /home/bmscecse/nltk_data...
nltk_data]   Package stopwords is already up-to-date!
5/05/20 12:27:48 WARN Utils: Your hostname, bmscecse-HP-Elite-Tower-600-G9-Desktop-PC resolves to a loopback address: 127.0.1.1; using 10.124.3.72 instead (on interface eno1)
5/05/20 12:27:48 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
5/05/20 12:27:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
sing Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
etting default log level to "WARN".
o adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
-------------------------------------------                                   (0 + 1) / 1]
ime: 2025-05-20 12:27:52
-------------------------------------------
```

bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~/spark-streaming-cleaner
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: ~/spark-streaming-cleaner

```
mscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ sudo apt install netcat
[sudo] password for bmscecse:
eading package lists... Done
uilding dependency tree... Done
eading state information... Done
he following NEW packages will be installed:
  netcat
 upgraded, 1 newly installed, 0 to remove and 614 not upgraded.
eed to get 2,044 B of archives.
fter this operation, 17.4 kB of additional disk space will be used.
et:1 http://in.archive.ubuntu.com/ubuntu jammy/universe amd64 netcat all 1.218-4ubuntu1 [2,044 B]
etched 2,044 B in 1s (3,758 B/s)
electing previously unselected package netcat.
(Reading database ... 172726 files and directories currently installed.)
reparing to unpack .../netcat_1.218-4ubuntu1_all.deb ...
npacking netcat (1.218-4ubuntu1) ...
etting up netcat (1.218-4ubuntu1) ...
mscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ nc -lk 9999
 Hello! Spark is cool and useful.
 We are learning Spark Streaming in Ubuntu.
```

```
Time: 2025-05-20 12:28:06
-------------------------------------------

-------------------------------------------
Time: 2025-05-20 12:28:08
-------------------------------------------

-------------------------------------------
Time: 2025-05-20 12:28:10
-------------------------------------------

-------------------------------------------
Time: 2025-05-20 12:28:12
-------------------------------------------

-------------------------------------------
Time: 2025-05-20 12:28:14
-------------------------------------------
hello
spark
cool
useful
learning
spark
streaming
ubuntu

-------------------------------------------
```

Scala prog 1 and 3https://chatgpt.com/share/682c2c05-1ce4-800b-9c13-a107fd371324Scala prog 2 https://chatgpt.com/share/682c2c14-0fe4-800b-94a5-2fcd2bd94c91