EXERCISE NO.: 03
WORD COUNT PROGRAM USING MAPREDUCE

AIM:

To implement a MapReduce program in Hadoop that counts the frequency of words in a text file, thereby demonstrating distributed data processing and the working of the MapReduce framework.

SCRIPT:

sample.txt

```
Hadoop is an open-source framework that allows for the distributed processing of
large data sets across clusters of computers using simple programming models
```

mapper.py

```
import sys
for line in sys.stdin:
    for word in line.strip().split():
        print(f"{word}\\t1")
```

reducer.py

```
import sys
from collections import defaultdict
word_count = defaultdict(int)
for line in sys.stdin:
    word, count = line.strip().split('\\t')
    word_count[word] += int(count)
for word, count in word_count.items():
    print(f"{word}\\t{count}")
```

Upload the file to HDFS

```
!hdfs dfs -mkdir -p /user/bdt/wordcount/input
!hdfs dfs -put input.txt /user/bdt/wordcount/input/
```

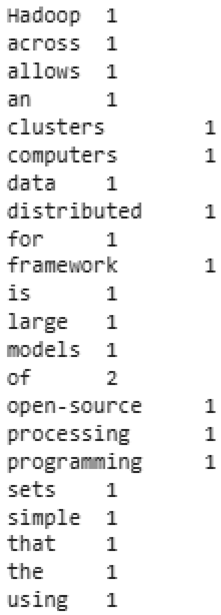
MapReduce Job: Hadoop Streaming

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar \
 -input /user/bdt/wordcount/input/ \
 -output /user/bdt/wordcount/output/ \
 -mapper mapper.py \
 -reducer reducer.py \
 -file mapper.py \
 -file reducer.py
```

<u>View result:</u>

```
!hdfs dfs -cat /user/colab/wordcount/output/part-00000
```

<u>OUTPUT:</u>



```
Hadoop  1
across  1
allows  1
an      1
clusters        1
computers       1
data    1
distributed     1
for     1
framework       1
is      1
large   1
models  1
of      2
open-source     1
processing      1
programming     1
sets    1
simple  1
that    1
the     1
using   1
```

<u>RESULT:</u>

Thus, the Word Count program using MapReduce was successfully implemented.