**EXP 2: Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm.**

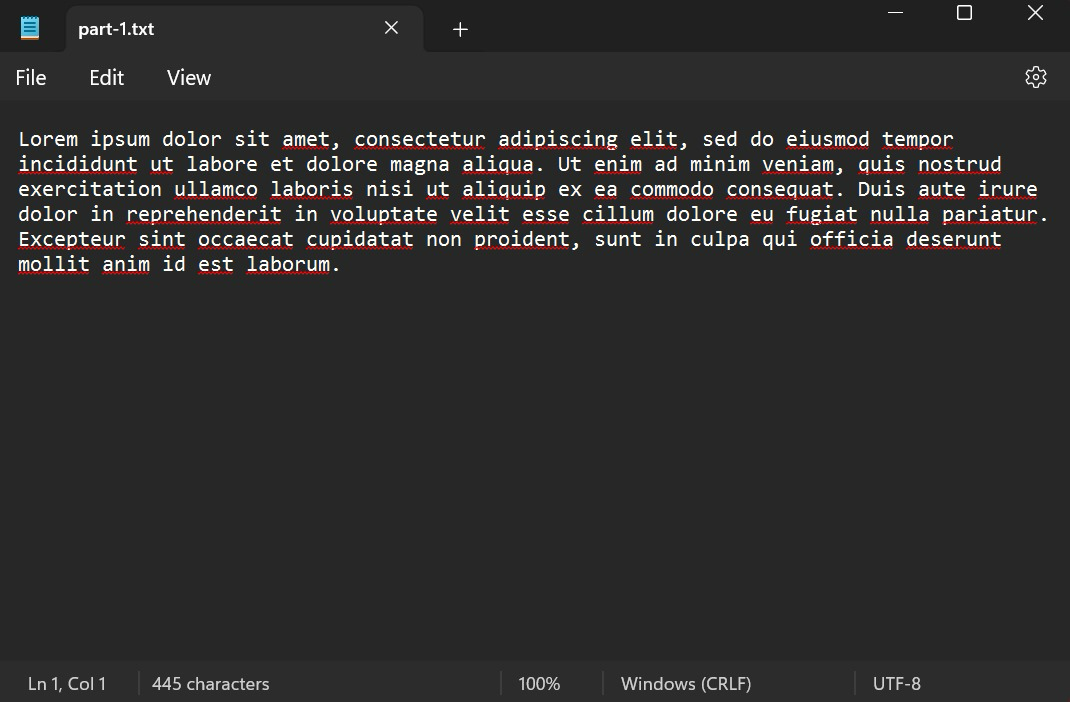
**AIM:**

To run a basic Word Count MapReduce program using Hadoop.

**PROCEDURE:**

**Step 1: Create Data File:**

Create a file named "part-1.txt" and populate it with text data that you wish to analyse.



**Step 2: Mapper Logic - mapper.py:**

Create a file named "mapper.py" to implement the logic for the mapper. The mapper will read input data from STDIN, split lines into words, and output each word with its count.

**mapper.py:**

#!C:/Users/user/AppData/Local/Microsoft/WindowsApps/python.exe

import sys

for line in sys.stdin:

line = line.strip()

words = line.split()

for word in words:

print('%s\t%s'%(word,1))

**Step 3: Reducer Logic - reducer.py:**

Create a file named "reducer.py" to implement the logic for the reducer. The reducer will aggregate the occurrences of each word and generate the final output.

**reducer.py:**

#!C:/Users/user/AppData/Local/Microsoft/WindowsApps/python.exe

import sys

prev\_word = None

prev\_count = 0

for line in sys.stdin:

line = line.strip()

word, count = line.split('\t')

count = int(count)

if prev\_word == word:

prev\_count += count

else:

if prev\_word:

print('%s\t%s' %(prev\_word, prev\_count))

prev\_count = count

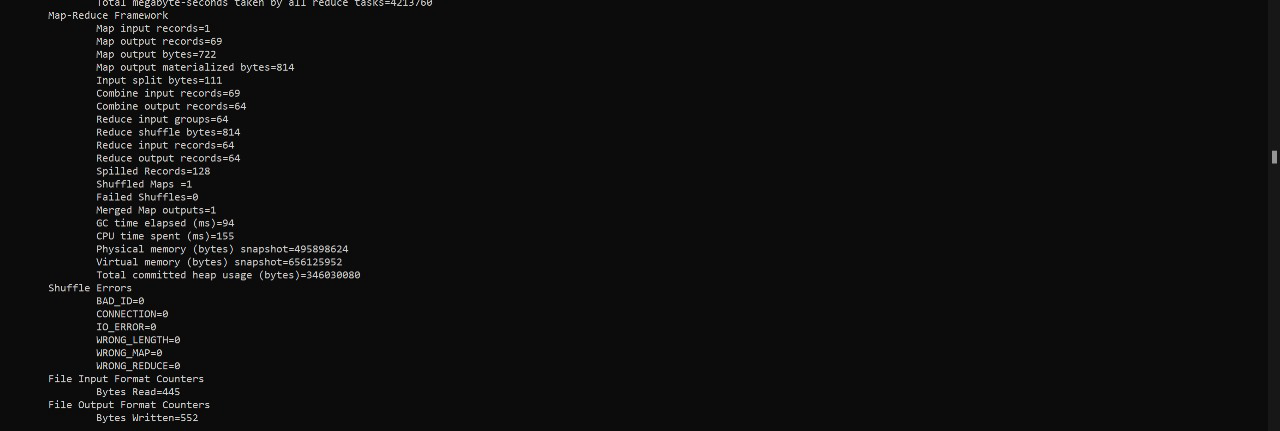
prev\_word = word

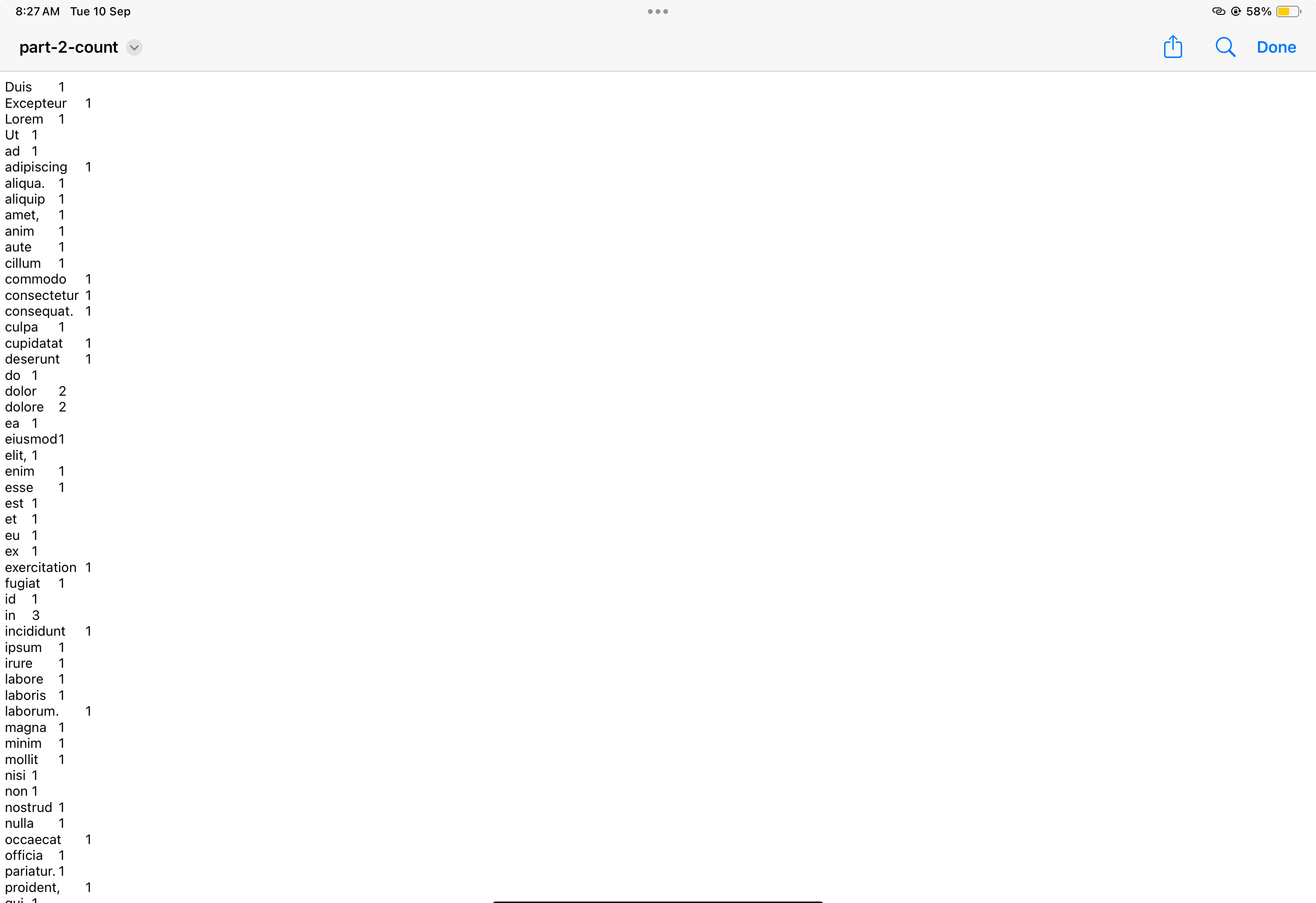
if prev\_word == word:

print('%s\t%s' %(prev\_word, prev\_count))

**Step 4: Prepare Hadoop Environment:**

|  |  |
| --- | --- |
| Start the Hadoop daemons and create a directory in HDFS to store your data. Run the following commands to store the data in the WordCount Directory.  *start-all.cmd*  *cd C:/Hadoop/sbin*  *hdfs dfs -mkdir /test*  *hdfs dfs -put C:/Users/user/Documents/DataAnalytics/input.txt /WordCount*  *hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-2.7.0.jar ^*  *-input /WordCount/input.txt ^*  *-output /WordCount/output ^*  *-mapper "python C:/ Users/user/Documents/DataAnalytics/mapper.py" ^*  *-reducer "python C:/ Users/user/Documents/DataAnalytics/reducer.py"*  **Step 5: Check Output:**   |  | | --- | | Check the output of the Word Count program in the specified HDFS output directory.  *hdfs dfs -cat /WordCount/output/part-00000*  **OUTPUT:** | |





**RESULT:**

Thus, the program for basic Word Count Map Reduce has been executed successfully.