Problem 2:

Source Code in PySpark:

import pyspark

from pyspark import SparkContext

from pyspark.mllib.linalg.distributed import CoordinateMatrix, MatrixEntry

from operator import add

from pyspark.sql import SparkSession

sc = SparkContext()

r=sc.textFile("part-00000")

m=r.flatMap(lambda x: x.split('\n')).filter(lambda x : "A" in x).map(lambda x : (x.strip("A, ")).split(' ')).map(lambda x: tuple(list(map(int, x))))

spark = SparkSession(sc)

print(hasattr(m,"toDF"))

cmat=CoordinateMatrix(m)

print(cmat.numRows()) # 3

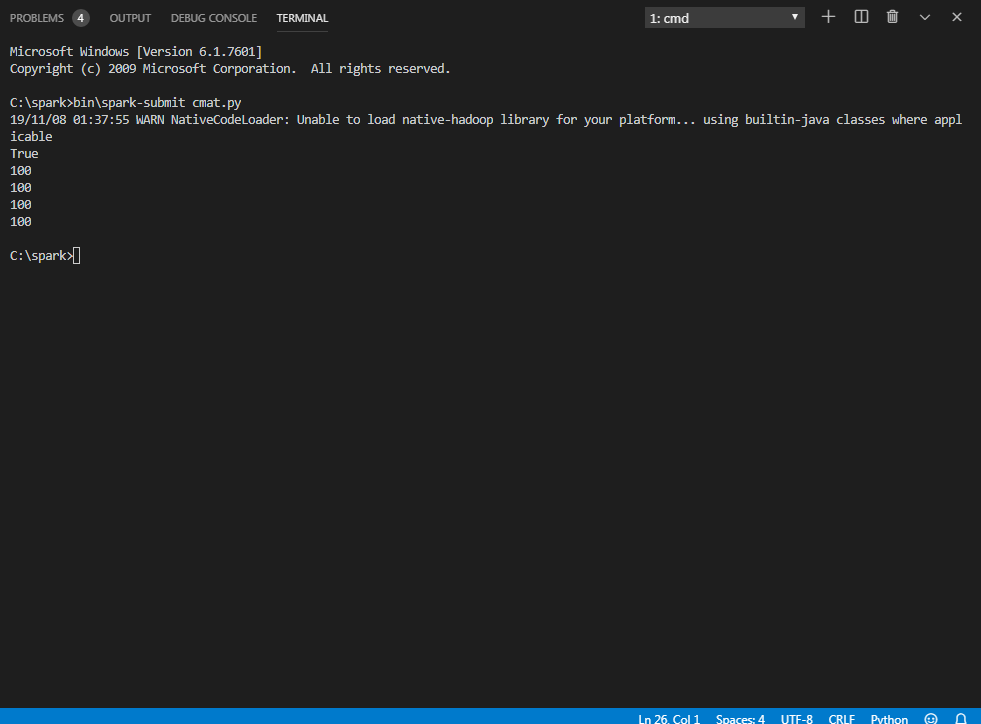
print(cmat.numCols())

rowmat = cmat.toRowMatrix()

print(rowmat.numRows()) # 3

print(rowmat.numCols())

Output:



Comments:

In this pyspark program , we have imported from the mllib i.e. machine learning library of spark various modules like Coordinate Matrix etc. to work with the matrices in spark and other modules like Spark Session and Spark Context for normal working of the program.

Important Steps

* Took input from the input file and made it into an RDD using sc.textfile.
* Extracted the A matrix in the coordinate format and then converted each coordinate with value into (Row, Column , Value) Tuples
* Passed this rdd of tuples to the CoordinateMatrix function to convert it into Coordinate Matrix RDD used in Pyspark.
* Converted this Coordinate Matrix RDD into Row matrix RDD using the toRowMatrix function on the Corrdinate Matrix RDD.
* Printed the Number of Rows and Number of columns attributes of both Matrices formats RDDs to show the successful formation of those RDDs.