Data: EntrKeyWords *(which contains a list of entrepreneurial keywords generated from entrepreneurship research journals)*

Perform the following tasks:

1. Write a MapReduce program so it generates the frequency of keywords listed in the EntrKeyWords.

```
[hadoop@ip-172-31-30-82 ~]$ nano mapper.py
[hadoop@ip-172-31-30-82 ~]$ nano reducer.py
[hadoop@ip-172-31-30-82 ~]$ chmod +x mapper.py
[hadoop@ip-172-31-30-82 ~]$ chmod +x reducer.py
[hadoop@ip-172-31-30-82 ~]$
```

```
[hadoop@ip-172-31-30-82 ~]$ find /usr/lib/ -name '*hadoop*streaming*.jar'
/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-2.jar
/usr/lib/hadoop/hadoop-streaming.jar
/usr/lib/hadoop-mapreduce/hadoop-streaming-3.3.6-amzn-2.jar
/usr/lib/hadoop-mapreduce/hadoop-streaming.jar
[hadoop@ip-172-31-30-82 ~]$
```

2. Create directories on the cluster and name these directories Entr2010 and Entr2020.

```
[hadoop@ip-172-31-30-82 ~]$ ls
Entr2010.txt  Entr2020.txt  EntrKeyWords.txt  NYSE.csv  mapper.py  reducer.py
[hadoop@ip-172-31-30-82 ~]$ hadoop fs -mkdir Entr2010
[hadoop@ip-172-31-30-82 ~]$ hadoop fs -mkdir Entr2020
[hadoop@ip-172-31-30-82 ~]$
```

3. Load the files from Entr2010 and Entr2020 in these new directories, respectively.

```
[hadoop@ip-172-31-22-61 ~]$ hadoop fs -mkdir Entr2010
^[[A[hadoop@ip-172-31-22-61 ~]$ hadoop fs -mkdir Entr2020
^[[A[hadoop@ip-172-31-22-61 ~]$ hadoop fs -copyFromLocal Entr2010.txt Entr2010/Entr2010.txt
[hadoop@ip-172-31-22-61 ~]$ hadoop fs -copyFromLocal Entr2020.txt Entr2020/Entr2020.txt
[hadoop@ip-172-31-22-61 ~]$
```

4. Run the MapReduce JAR twice, once for 2010, and once for 2020.

```
[hadoop@ip-172-31-22-61 ~]$ hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -files mapper.py,reducer.py -
mapper mapper.py -reducer reducer.py -input Entr2010/Entr2010.txt -output Entr2010/output
```

```
2024-03-13 21:32:36,102 INFO streaming.StreamJob: Output directory: Entr2010/output
[hadoop@ip-172-31-22-61 ~]$ hadoop fs -ls Entr2010/output
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2024-03-13 21:32 Entr2010/output/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup       2110 2024-03-13 21:32 Entr2010/output/part-00000
-rw-r--r--   1 hadoop hdfsadmingroup       2148 2024-03-13 21:32 Entr2010/output/part-00001
-rw-r--r--   1 hadoop hdfsadmingroup       2235 2024-03-13 21:32 Entr2010/output/part-00002
[hadoop@ip-172-31-22-61 ~]$
```

```
[hadoop@ip-172-31-22-61 ~]$ hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -files mapper.py,reducer.py -
mapper mapper.py -reducer reducer.py -input Entr2020/Entr2020.txt -output Entr2020/output
```

```
[hadoop@ip-172-31-22-61 ~]$ hadoop fs -ls Entr2020/output
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2024-03-13 22:17 Entr2020/output/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup        170 2024-03-13 22:17 Entr2020/output/part-00000
-rw-r--r--   1 hadoop hdfsadmingroup        144 2024-03-13 22:17 Entr2020/output/part-00001
-rw-r--r--   1 hadoop hdfsadmingroup        176 2024-03-13 22:17 Entr2020/output/part-00002
[hadoop@ip-172-31-22-61 ~]$ 3#
```

5. Observe and compare the frequencies of the keywords from 2010 and 2020.

```
[hadoop@ip-172-31-22-61 ~]$ hadoop fs -cat Entr2010/output1/p*
article 1
available       1
because 1
important       1
time    1
community       1
entrepreneurship        2
feeling 1
great   3
business        9
companies       5
entrepreneur    1
means   1
need    2
people  4
```

```
[hadoop@ip-172-31-22-61 ~]$ hadoop fs -cat Entr2020/output/p*
article 6
available       2
because 6
better  3
bored   7
create  3
culture 4
employees       3
enough  2
experience      2
important       2
life    4
progress        2
registry        2
time    4
workplace       2
years   5
career  3
choose  4
community       2
effective       3
entrepreneurship        3
episode 4
feeling 2
forward 4
great   3
helps   3
information     2
perseverance    1
world   3
before  4
bizarre 4
business        2
companies       4
entrepreneur    5
feel    4
generous        2
means   3
need    10
own     3
people  15
read    2
remarkable      5
resilient       2
selfish 4
talking 3
work    6
working 3
[hadoop@ip-172-31-22-61 ~]$
```