Question 1 (Midterm exam)

Name - Mihir Chhatre
NetID - mc9164
Environment used - Dataproc

Course - Big Data
Section - D
Semester - 3rd

Note: Please zoom-in on screenshots to see commands clearly.

Step 1: Uploaded all files from the cookbook directory and moved all of these new text files to a directory called cookbook.

```
mc9164_nyu_edu@nyu-dataproc-m:~$ mkdir cookbook
mc9164_nyu_edu@nyu-dataproc-m:~$ mv *.txt cookbook/
mc9164_nyu_edu@nyu-dataproc-m:~$
```

Step 2: Moved the cookbook directory from dataproc local to HDFS

```
mc9164_nyu_edu@nyu-dataproc-m:~$
mc9164_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put cookbook
mc9164_nyu_edu@nyu-dataproc-m:~$
```

Step 3: Running the MapReduce command with 700 as sample parameter

```
hadoop jar $HADOOP_HOME/hadoop-streaming-3.2.2.jar \
-input cookbook/*.txt \
-output Q1_result \
-mapper "python3 mapper_mid_q1.py" \
-reducer "python3 reducer_mid_q1.py 700" \
-file mapper_mid_q1.py \
-file reducer_mid_q1.py \
-numReduceTasks 1
```

```
mc9164_nyu_edu@nyu-dataproc-m:~$ hadoop jar $HADOOP_HOME/hadoop-streaming-3.2.2.jar \
> -input cookbook/*.txt \
> -output Q1_result \
> -mapper "python3 mapper_mid_q1.py" \
> -reducer "python3 reducer_mid_q1.py 700" \
> -file mapper_mid_q1.py \
> -file reducer_mid_q1.py \
> -numReduceTasks 1
```

Step 5: Checking is output file is generated correctly