Name: Ajinkya Vijay Sonawane | **NetID:** avs8687

NYU CS6513 Big Data Spring 2022
Assignment 1 - Hadoop HDFS & MapReduce

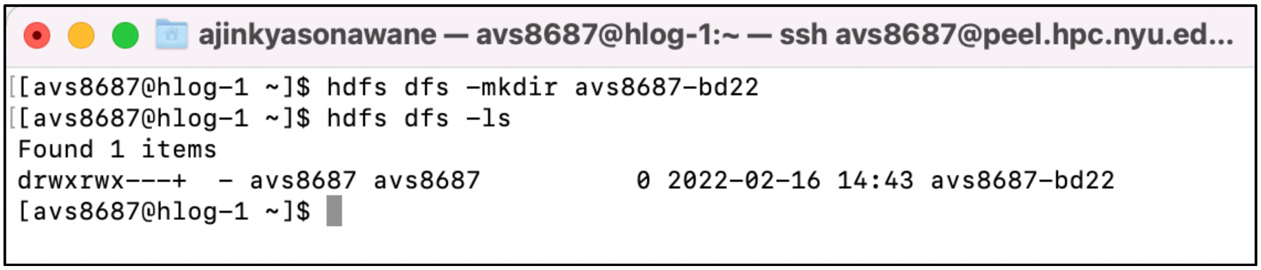
1. HDFS

Running any version of Hadoop (HPC, docker or otherwise), submit screen grabs (a picture in jpg or other suitable format) of the following:

a. create a directory in HDFS with this format: netid-bd22 (e.g. mine will be 'jcr365-bd22'). Submit a screen grab of the output of a Hadoop file listing showing your home directory and your new directory in it.

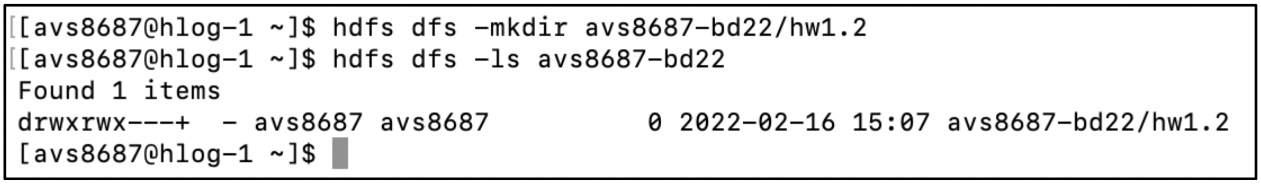
Commands to create directory:
hadoop fs -mkdir <DIRECTORY_NAME>
hdfs dfs -mkdir <DIRECTORY_NAME>

Commands to list directories:
hadoop fs -ls
hdfs dfs -ls


ajinkyasonawane — avs8687@hlog-1:~ — ssh avs8687@peel.hpc.nyu.ed...

```
[[avs8687@hlog-1 ~]$ hdfs dfs -mkdir avs8687-bd22
[[avs8687@hlog-1 ~]$ hdfs dfs -ls
Found 1 items
drwxrwx---+   - avs8687 avs8687          0 2022-02-16 14:43 avs8687-bd22
[avs8687@hlog-1 ~]$
```

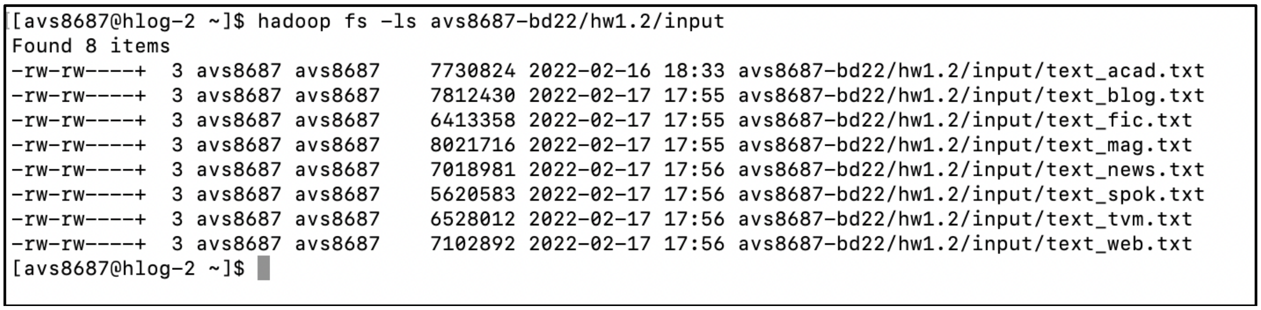
b. Create a directory for the homework problem 1.2 (bigram count), and extract all input files into it. Call this directory as follows: hw1.2, e.g. mine will be hw1.2. Submit a picture of directory listings or otherwise show the input files in it.



```
[[avs8687@hlog-1 ~]$ hdfs dfs -mkdir avs8687-bd22/hw1.2
[[avs8687@hlog-1 ~]$ hdfs dfs -ls avs8687-bd22
Found 1 items
drwxrwx---+   - avs8687 avs8687          0 2022-02-16 15:07 avs8687-bd22/hw1.2
[avs8687@hlog-1 ~]$
```

Name: Ajinkya Vijay Sonawane | **NetID:** avs8687

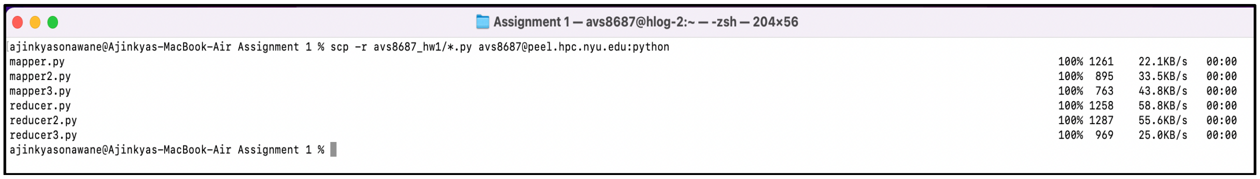
Upload files to HDFS: `hadoop fs -put hw1/*.txt avs8687-bd22/hw1.2/input`



```
[avs8687@hlog-2 ~]$ hadoop fs -ls avs8687-bd22/hw1.2/input
Found 8 items
-rw-rw----+  3 avs8687 avs8687    7730824 2022-02-16 18:33 avs8687-bd22/hw1.2/input/text_acad.txt
-rw-rw----+  3 avs8687 avs8687    7812430 2022-02-17 17:55 avs8687-bd22/hw1.2/input/text_blog.txt
-rw-rw----+  3 avs8687 avs8687    6413358 2022-02-17 17:55 avs8687-bd22/hw1.2/input/text_fic.txt
-rw-rw----+  3 avs8687 avs8687    8021716 2022-02-17 17:55 avs8687-bd22/hw1.2/input/text_mag.txt
-rw-rw----+  3 avs8687 avs8687    7018981 2022-02-17 17:56 avs8687-bd22/hw1.2/input/text_news.txt
-rw-rw----+  3 avs8687 avs8687    5620583 2022-02-17 17:56 avs8687-bd22/hw1.2/input/text_spok.txt
-rw-rw----+  3 avs8687 avs8687    6528012 2022-02-17 17:56 avs8687-bd22/hw1.2/input/text_tvm.txt
-rw-rw----+  3 avs8687 avs8687    7102892 2022-02-17 17:56 avs8687-bd22/hw1.2/input/text_web.txt
[avs8687@hlog-2 ~]$
```

2. Language Models with MapReduce

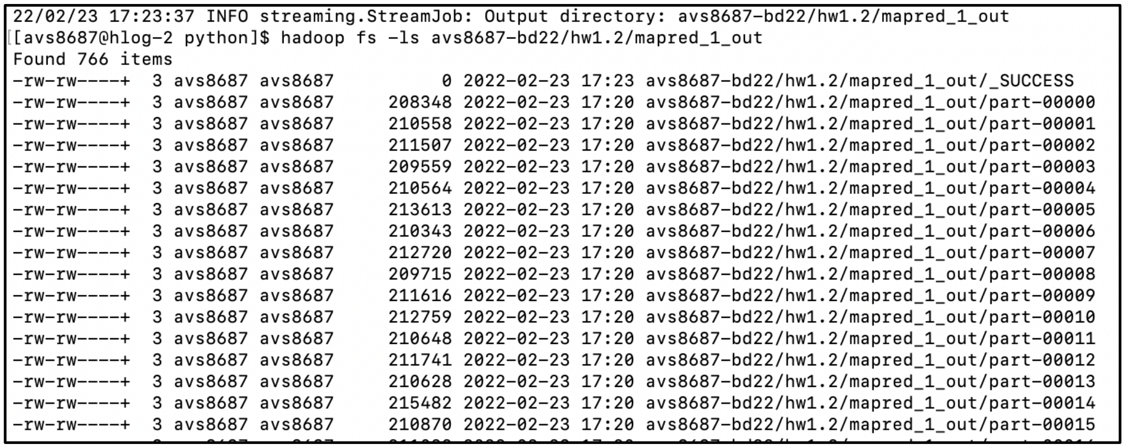
Upload Python files to Login Node using "scp":



```
ajinkyasonawane@Ajinkyas-MacBook-Air Assignment 1 % scp -r avs8687_hw1/*.py avs8687@peel.hpc.nyu.edu:python
mapper.py                                                                              100% 1261    22.1KB/s   00:00
mapper2.py                                                                             100%  895    33.5KB/s   00:00
mapper3.py                                                                             100%  763    43.8KB/s   00:00
reducer.py                                                                             100% 1258    58.8KB/s   00:00
reducer2.py                                                                            100% 1287    55.6KB/s   00:00
reducer3.py                                                                            100%  969    25.0KB/s   00:00
ajinkyasonawane@Ajinkyas-MacBook-Air Assignment 1 %
```

Run the First Map Reduce Operation to map n-gram and '1' to count their frequency in the reducer:

```
mapred streaming --files mapper.py,reducer.py -mapper "python
mapper.py" -reducer "python reducer.py" -input avs8687-
bd22/hw1.2/input -output avs8687-bd22/hw1.2/mapred_1_out
```

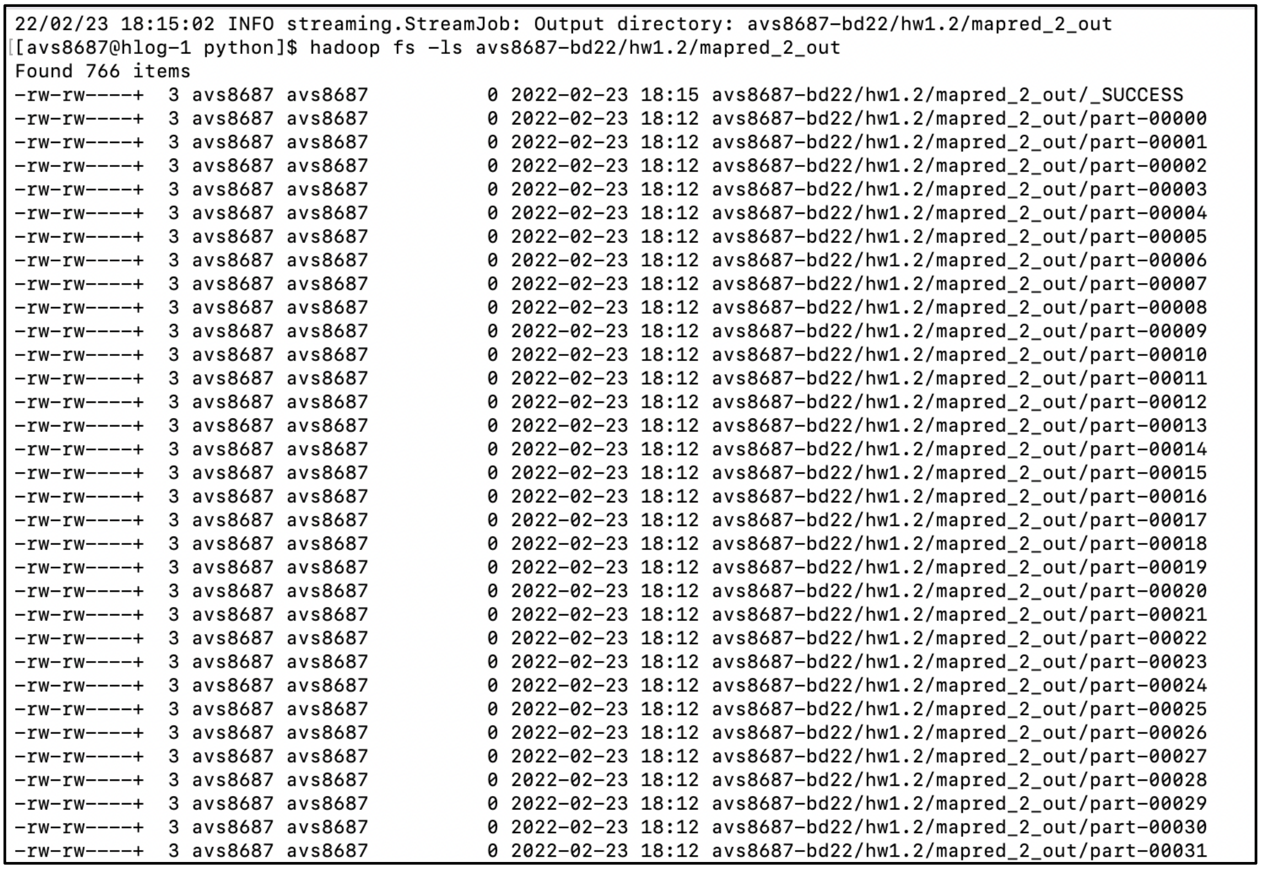


```
22/02/23 17:23:37 INFO streaming.StreamJob: Output directory: avs8687-bd22/hw1.2/mapred_1_out
[avs8687@hlog-2 python]$ hadoop fs -ls avs8687-bd22/hw1.2/mapred_1_out
Found 766 items
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 17:23 avs8687-bd22/hw1.2/mapred_1_out/_SUCCESS
-rw-rw----+  3 avs8687 avs8687     208348 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00000
-rw-rw----+  3 avs8687 avs8687     210558 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00001
-rw-rw----+  3 avs8687 avs8687     211507 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00002
-rw-rw----+  3 avs8687 avs8687     209559 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00003
-rw-rw----+  3 avs8687 avs8687     210564 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00004
-rw-rw----+  3 avs8687 avs8687     213613 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00005
-rw-rw----+  3 avs8687 avs8687     210343 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00006
-rw-rw----+  3 avs8687 avs8687     212720 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00007
-rw-rw----+  3 avs8687 avs8687     209715 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00008
-rw-rw----+  3 avs8687 avs8687     211616 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00009
-rw-rw----+  3 avs8687 avs8687     212759 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00010
-rw-rw----+  3 avs8687 avs8687     210648 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00011
-rw-rw----+  3 avs8687 avs8687     211741 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00012
-rw-rw----+  3 avs8687 avs8687     210628 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00013
-rw-rw----+  3 avs8687 avs8687     215482 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00014
-rw-rw----+  3 avs8687 avs8687     210870 2022-02-23 17:20 avs8687-bd22/hw1.2/mapred_1_out/part-00015
```

Name: Ajinkya Vijay Sonawane | **NetID:** avs8687

Run the Second Map Reduce Operation to map n-grams to their type:
(1-unigram, 2-bigram, 3-trigram)

```
mapred streaming --files mapper2.py,reducer2.py -mapper "python
mapper2.py" -reducer "python reducer2.py" -input avs8687-
bd22/hw1.2/mapred_1_out -output avs8687-bd22/hw1.2/mapred_2_out
```

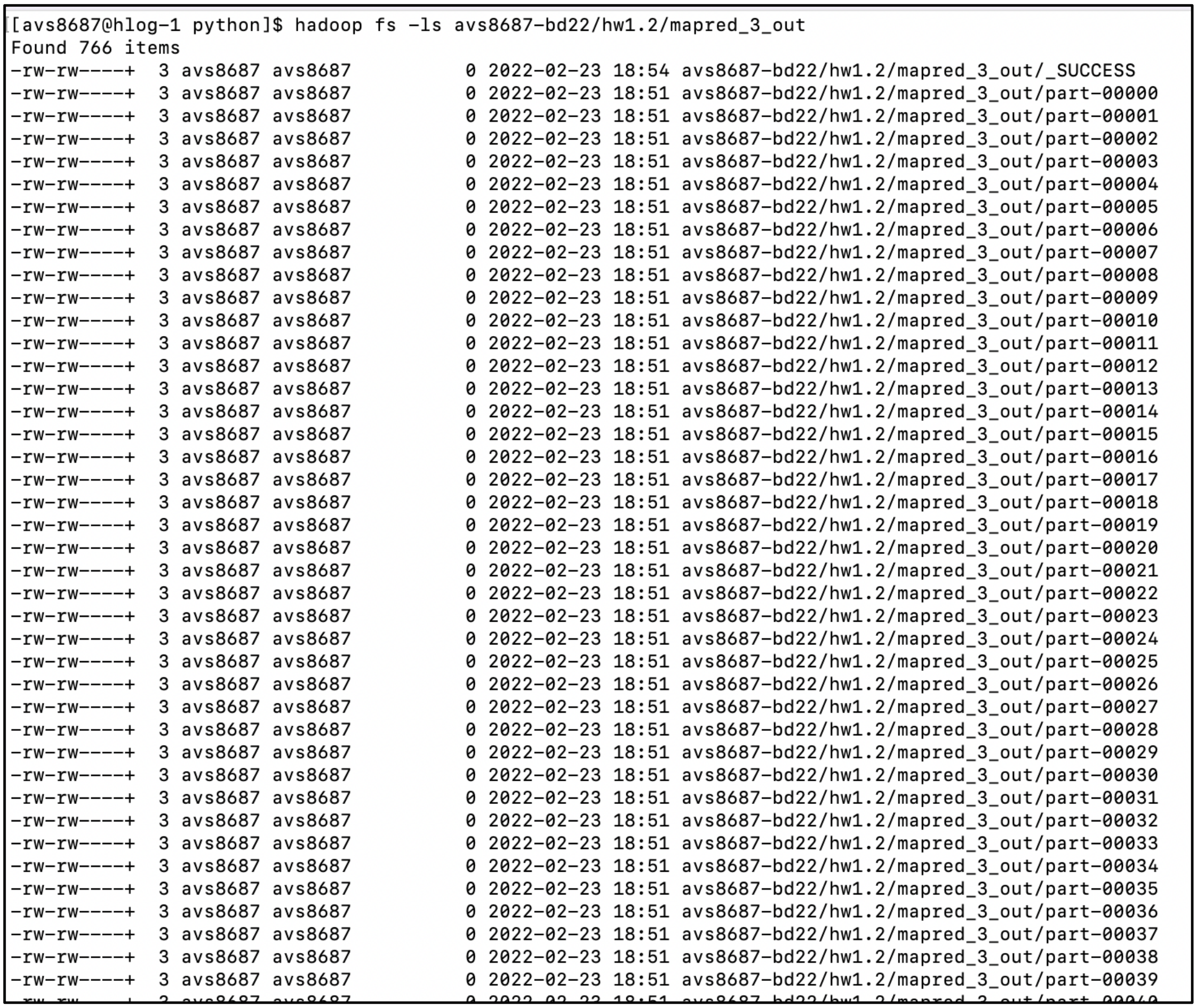


```
22/02/23 18:15:02 INFO streaming.StreamJob: Output directory: avs8687-bd22/hw1.2/mapred_2_out
[[avs8687@hlog-1 python]$ hadoop fs -ls avs8687-bd22/hw1.2/mapred_2_out
Found 766 items
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:15 avs8687-bd22/hw1.2/mapred_2_out/_SUCCESS
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00000
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00001
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00002
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00003
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00004
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00005
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00006
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00007
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00008
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00009
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00010
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00011
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00012
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00013
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00014
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00015
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00016
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00017
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00018
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00019
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00020
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00021
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00022
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00023
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00024
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00025
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00026
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00027
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00028
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00029
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00030
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:12 avs8687-bd22/hw1.2/mapred_2_out/part-00031
```

3. Extra Credit

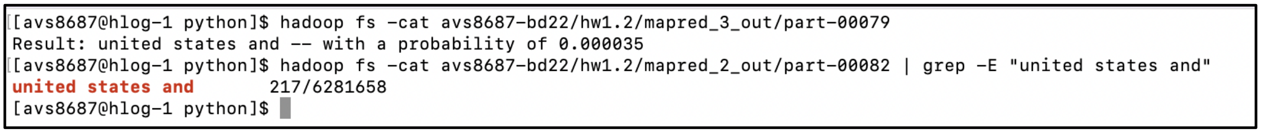
Run the 3rd Map Reduce operation to map all trigrams containing “united states” and find the trigram having the highest probability

```
mapred streaming --files mapper3.py,reducer3.py -mapper "python
mapper3.py" -reducer "python reducer3.py" -input avs8687-
bd22/hw1.2/mapred_2_out -output avs8687-bd22/hw1.2/mapred_3_out
```



```
[avs8687@hlog-1 python]$ hadoop fs -ls avs8687-bd22/hw1.2/mapred_3_out
Found 766 items
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:54 avs8687-bd22/hw1.2/mapred_3_out/_SUCCESS
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00000
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00001
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00002
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00003
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00004
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00005
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00006
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00007
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00008
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00009
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00010
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00011
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00012
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00013
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00014
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00015
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00016
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00017
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00018
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00019
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00020
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00021
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00022
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00023
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00024
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00025
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00026
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00027
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00028
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00029
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00030
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00031
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00032
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00033
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00034
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00035
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00036
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00037
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00038
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00039
-rw-rw----+  3 avs8687 avs8687          0 2022-02-23 18:51 avs8687-bd22/hw1.2/mapred_3_out/part-00040
```

Final Answer:



```
[avs8687@hlog-1 python]$ hadoop fs -cat avs8687-bd22/hw1.2/mapred_3_out/part-00079
Result: united states and -- with a probability of 0.000035
[avs8687@hlog-1 python]$ hadoop fs -cat avs8687-bd22/hw1.2/mapred_2_out/part-00082 | grep -E "united states and"
united states and       217/6281658
[avs8687@hlog-1 python]$
```