```
(base) pragatwagle@Pragats-MacBook-Air hadoop-3.3.4 % hadoop jar \
     /Users/pragatwagle/hadoop-3.3.4/share/hadoop/tools/sources/hadoop-streaming-3.3.4.jar \
     -files /Users/pragatwagle/Desktop/Hadoop/HW1/mapper.py,/Users/pragatwagle/Desktop/Hadoop/HW1/reducer.py \
     -mapper mapper.py -reducer reducer.py \
     -input /hwone/iub_wiki.txt -output /hwone/output
2022-09-11 15:00:22,759 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/var/folders/lj/chp0k_rn1qqf1qs2mtg65xzw0000gn/T/hadoop-unjar6088084093744009598/] [] /var/folders/lj/chp0k_rn1qqf1qs2mtg65xzw0000gn/T/stre
amjob4863509746635531037.jar tmpDir=null
2022-09-11 15:00:23,101 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-09-11 15:00:23,191 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-09-11 15:00:23,357 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/pragatwagle/.staging/job_1662922192
035_0003
2022-09-11 15:00:24,074 INFO mapred.FileInputFormat: Total input files to process : 1
2022-09-11 15:00:24,806 INFO mapreduce.JobSubmitter: number of splits:2
2022-09-11 15:00:25,465 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1662922192035_0003
2022-09-11 15:00:25,466 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-09-11 15:00:25,592 INFO conf.Configuration: resource-types.xml not found
2022-09-11 15:00:25,592 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-09-11 15:00:25,616 INFO impl.YarnClientImpl: Submitted application application_1662922192035_0003
2022-09-11 15:00:25,629 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1662922192035_0003/
2022-09-11 15:00:25,630 INFO mapreduce.Job: Running job: job_1662922192035_0003
2022-09-11 15:00:33,847 INFO mapreduce.Job: Job job_1662922192035_0003 running in uber mode : false
2022-09-11 15:00:33,850 INFO mapreduce.Job:  map 0% reduce 0%
2022-09-11 15:00:39,961 INFO mapreduce.Job:  map 100% reduce 0%
```

```
2022-09-11 15:00:50,248 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=6279
                FILE: Number of bytes written=851683
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3843
                HDFS: Number of bytes written=7879
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=8928
                Total time spent by all reduces in occupied slots (ms)=4866
                Total time spent by all map tasks (ms)=8928
                Total time spent by all reduce tasks (ms)=4866
                Total vcore-milliseconds taken by all map tasks=8928
                Total vcore-milliseconds taken by all reduce tasks=4866
                Total megabyte-milliseconds taken by all map tasks=9142272
```

```
                Reduce input records=395
                Reduce output records=187
                Spilled Records=790
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=48
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=614465536
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=3669
        File Output Format Counters
                Bytes Written=7879
2022-09-11 15:00:50,248 INFO streaming.StreamJob: Output directory: /hwone/output
(base) pragatwagle@Pragats-MacBook-Air hadoop-3.3.4 %
```

```
WORD tab {Document1:[Count, [WordIndexs]]...DocumentN:[Count,
[WordIndexs]]}
all     {'doc4': [1, ['54']]}
athletic{'doc5': [1, ['71']]}
over    {'doc1': [1, ['30']]}
global  {'doc3': [1, ['70']]}
schools {'doc3': [1, ['14']]}
including       {'doc3': [1, ['17']]}
go      {'doc4': [1, ['76']]}
its     {'doc1': [1, ['46']]}
jacobs  {'doc3': [1, ['19']]}
based   {'doc2': [1, ['41']]}
located {'doc1': [1, ['22']]}
with    {'doc2': [1, ['36']], 'doc1': [1, ['29']], 'doc5': [2, ['58',
'49']]}
institution     {'doc1': [1, ['39']]}
to      {'doc1': [1, ['11']], 'doc5': [1, ['43']]}
program {'doc2': [1, ['35']], 'doc5': [1, ['48']]}
systems {'doc2': [1, ['74']]}
2015    {'doc5': [1, ['4']]}
8       {'doc3': [1, ['74']]}
extensive       {'doc5': [1, ['45']]}
has     {'doc2': [1, ['33']], 'doc3': [1, ['12']]}
division{'doc4': [1, ['40']], 'doc5': [1, ['75']]}
executive       {'doc2': [1, ['67']]}
than    {'doc4': [2, ['72', '64']], 'doc5': [2, ['51', '60']]}
organizations   {'doc4': [1, ['67']], 'doc5': [1, ['54']]}
ceo     {'doc2': [1, ['71']]}
big     {'doc4': [1, ['42']], 'doc5': [1, ['94']]}
largest {'doc1': [1, ['47']]}
academics       {'doc4': [1, ['4']]}
were    {'doc5': [1, ['35']]}
undergraduates  {'doc5': [1, ['64']]}
fall    {'doc5': [1, ['3']]}
not     {'doc4': [1, ['50']]}
colloquially    {'doc1': [1, ['9']]}
association     {'doc3': [1, ['7']]}
carmichael      {'doc2': [1, ['58']]}
school  {'doc2': [5, ['12', '18', '26', '15', '32']], 'doc3': [11,
['31', '35', '24', '20', '40', '44', '48', '56', '60', '65', '68']],
'doc4': [2, ['80', '30']]}
d       {'doc5': [1, ['29']]}
level   {'doc2': [1, ['7']]}
55      {'doc5': [1, ['12']]}
numerous{'doc3': [1, ['13']]}
university      {'doc1': [4, ['48', '43', '2', '21']], 'doc4': [1,
['48']], 'doc5': [3, ['10', '40', '88']]}
50      {'doc4': [1, ['23']], 'doc5': [1, ['26']]}
teams   {'doc4': [2, ['33', '55']], 'doc5': [1, ['72']]}
large   {'doc4': [1, ['82']]}
```

```
team    {'doc2': [1, ['40']]}
notorious       {'doc4': [1, ['35']]}
referred{'doc1': [1, ['10']]}
abbreviated     {'doc1': [1, ['4']]}
cisco   {'doc2': [1, ['73']]}
ten     {'doc4': [1, ['43']], 'doc5': [1, ['95']]}
songwriter      {'doc2': [1, ['56']]}
national{'doc4': [1, ['14']]}
are     {'doc2': [1, ['9']], 'doc4': [3, ['34', '62', '56']], 'doc5':
[1, ['81']]}
year    {'doc2': [1, ['39']]}
chambers{'doc2': [1, ['66']]}
universities    {'doc3': [1, ['10']], 'doc4': [2, ['25', '15']]}
pictures{'doc2': [1, ['88']]}
magnolia{'doc2': [1, ['87']]}
since   {'doc4': [1, ['46']]}
legal   {'doc2': [1, ['48']]}
research{'doc1': [1, ['20']]}
include {'doc2': [1, ['53']]}
3       {'doc5': [1, ['38']]}
does    {'doc4': [1, ['49']]}
health  {'doc3': [1, ['38']]}
mavericks       {'doc2': [1, ['83']]}
7       {'doc1': [1, ['49']]}
international   {'doc3': [1, ['72']]}
approach{'doc2': [1, ['42']]}
public  {'doc2': [1, ['20']], 'doc3': [2, ['50', '37']], 'doc1': [1,
['19']], 'doc4': [1, ['24']]}
body    {'doc5': [1, ['17']]}
maurer  {'doc2': [1, ['25']], 'doc3': [1, ['55']]}
business{'doc2': [1, ['14']], 'doc3': [1, ['33']]}
of      {'doc2': [6, ['80', '72', '19', '16', '13', '27']], 'doc3':
[12, ['69', '57', '5', '8', '49', '45', '41', '21', '36', '25', '61',
'32']], 'doc1': [1, ['40']], 'doc4': [2, ['3', '84']], 'doc5': [6,
['77', '14', '2', '63', '92', '24']]}
notable {'doc2': [1, ['50']]}
informatics     {'doc3': [1, ['26']]}
100     {'doc4': [1, ['13']]}
on      {'doc2': [1, ['62']], 'doc4': [1, ['68']], 'doc5': [1,
['55']]}
c       {'doc5': [1, ['30']]}
country {'doc4': [1, ['28']]}
ncaa    {'doc4': [1, ['39']], 'doc5': [1, ['79']]}
foreign {'doc5': [1, ['33']]}
american{'doc3': [1, ['9']]}
studies {'doc3': [1, ['73']]}
514     {'doc5': [1, ['6']]}
first   {'doc2': [1, ['38']]}
among   {'doc2': [1, ['1']]}
there   {'doc4': [1, ['61']]}
```

```
washington      {'doc5': [1, ['28']]}
community       {'doc4': [1, ['83']]}
simply  {'doc1': [1, ['15']], 'doc4': [1, ['58']]}
composer{'doc2': [1, ['54']]}
owner   {'doc2': [1, ['79']]}
sports  {'doc4': [1, ['32']]}
hoagy   {'doc2': [1, ['57']]}
conference      {'doc4': [1, ['44']], 'doc5': [1, ['96']]}
from    {'doc2': [1, ['45']], 'doc5': [2, ['19', '22']]}
top     {'doc4': [2, ['22', '12']]}
sororities      {'doc4': [1, ['87']]}
system  {'doc1': [1, ['44']], 'doc5': [1, ['68']]}
mark    {'doc2': [1, ['76']]}
2       {'doc5': [1, ['13']]}
music   {'doc3': [1, ['22']]}
650     {'doc4': [1, ['65']]}
6       {'doc1': [1, ['28']]}
bloomington     {'doc2': [1, ['3']], 'doc1': [4, ['35', '24', '6',
'3']], 'doc4': [1, ['9']]}
john    {'doc2': [1, ['65']]}
was     {'doc5': [1, ['18']]}
more    {'doc4': [2, ['71', '63']], 'doc5': [2, ['59', '50']]}
life    {'doc5': [1, ['47']]}
750     {'doc5': [1, ['52']]}
mind    {'doc2': [1, ['64']]}
criteria{'doc4': [1, ['7']]}
environmental   {'doc2': [1, ['22']], 'doc3': [1, ['52']]}
penned  {'doc2': [1, ['60']]}
indiana {'doc2': [2, ['51', '29']], 'doc1': [4, ['1', '42', '25',
'16']], 'doc4': [1, ['47']], 'doc5': [4, ['85', '70', '20', '9']]}
known   {'doc4': [1, ['57']], 'doc5': [1, ['82']]}
165     {'doc5': [1, ['32']]}
fraternities    {'doc4': [1, ['85']]}
former  {'doc2': [1, ['70']]}
mascot  {'doc4': [1, ['53']]}
10      {'doc4': [1, ['88']]}
georgia {'doc2': [1, ['61']]}
17      {'doc5': [1, ['61']]}
nations {'doc5': [1, ['34']]}
flagship{'doc1': [1, ['38']]}
ranks   {'doc4': [1, ['10']]}
while   {'doc5': [1, ['11']]}
many    {'doc2': [1, ['5']]}
my      {'doc2': [1, ['63']]}
and     {'doc2': [6, ['69', '24', '86', '55', '75', '21']], 'doc3':
[6, ['71', '66', '11', '51', '15', '27']], 'doc1': [2, ['8', '45']],
'doc4': [5, ['70', '45', '86', '20', '5']], 'doc5': [3, ['80', '31',
'57']]}
nursing {'doc3': [1, ['42']]}
computing       {'doc3': [1, ['28']]}
```

```
is      {'doc3': [1, ['2']], 'doc1': [2, ['36', '17']], 'doc5': [2,
['89', '41']]}
in      {'doc1': [1, ['23']], 'doc4': [6, ['37', '26', '16', '1',
'11', '78']], 'doc5': [1, ['74']]}
iu      {'doc2': [1, ['2']], 'doc1': [3, ['34', '5', '13']], 'doc4':
[1, ['8']]}
it      {'doc3': [1, ['1']]}
an      {'doc5': [1, ['44']]}
states  {'doc1': [1, ['27']], 'doc4': [1, ['19']], 'doc5': [1,
['27']]}
as      {'doc1': [1, ['12']], 'doc4': [1, ['59']], 'doc5': [2, ['1',
'83']]}
have    {'doc4': [1, ['51']]}
home    {'doc5': [1, ['42']]}
education       {'doc2': [2, ['49', '17']], 'doc3': [1, ['62']]}
kelley  {'doc2': [1, ['11']], 'doc3': [1, ['30']]}
campus  {'doc4': [1, ['69']], 'doc5': [1, ['56']]}
affairs {'doc2': [1, ['23']], 'doc3': [1, ['53']]}
united  {'doc1': [1, ['26']], 'doc4': [1, ['18']]}
48      {'doc1': [1, ['31']], 'doc5': [1, ['5']]}
49      {'doc5': [1, ['23']]}
media   {'doc3': [1, ['64']]}
percent {'doc5': [1, ['62']]}
s       {'doc2': [2, ['4', '30']], 'doc4': [2, ['81', '31']]}
graduate{'doc2': [1, ['6']]}
member  {'doc3': [1, ['4']], 'doc5': [1, ['91']]}
also    {'doc5': [1, ['36']]}
other   {'doc4': [1, ['6']]}
5       {'doc1': [1, ['7']], 'doc4': [1, ['73']]}
chairman{'doc2': [1, ['68']]}
billionaire     {'doc2': [1, ['78']]}
9       {'doc5': [1, ['69']]}
optometry       {'doc3': [1, ['46']]}
attend  {'doc5': [1, ['8']]}
students{'doc1': [1, ['33']], 'doc4': [1, ['75']], 'doc5': [2, ['21',
'7']]}
who     {'doc2': [1, ['59']]}
theatres{'doc2': [1, ['85']]}
diversion       {'doc2': [1, ['44']]}
000     {'doc1': [1, ['32']], 'doc4': [1, ['74']]}
student {'doc4': [1, ['66']], 'doc5': [3, ['16', '53', '46']]}
law     {'doc2': [2, ['31', '28']], 'doc3': [1, ['58']]}
a       {'doc2': [3, ['43', '37', '34']], 'doc3': [1, ['3']], 'doc1':
[1, ['18']], 'doc4': [1, ['52']], 'doc5': [1, ['90']]}
cuban   {'doc2': [1, ['77']]}
hoosiers{'doc4': [1, ['60']], 'doc5': [1, ['86']]}
programs{'doc2': [1, ['8']], 'doc3': [1, ['16']]}
i       {'doc4': [1, ['41']], 'doc5': [1, ['76']]}
dallas  {'doc2': [1, ['82']]}
alumni  {'doc2': [1, ['52']]}
```

```
or      {'doc1': [1, ['14']]}
greek   {'doc4': [1, ['77']], 'doc5': [1, ['67']]}
competitors     {'doc4': [1, ['36']]}
compete {'doc5': [1, ['73']]}
enrolled{'doc5': [1, ['37']]}
landmark{'doc2': [1, ['84']]}
the     {'doc2': [3, ['10', '46', '81']], 'doc3': [12, ['67', '18',
'39', '6', '23', '54', '59', '63', '34', '43', '47', '29']], 'doc1':
[2, ['41', '37']], 'doc4': [6, ['17', '21', '79', '38', '27', '29']],
'doc5': [8, ['84', '87', '93', '66', '39', '78', '15', '25']]}
typical {'doc2': [1, ['47']]}
terms   {'doc4': [1, ['2']]}
joining {'doc5': [1, ['65']]}
```

Explanation:

In this case each paragraph was considered to be a different document new line characters were introduced between each paragraph. The mapper function reads from stdin which is the iub_wiki.txt file and for each line in the file creates a new document id and strips and finds all words in the line and then prints out the word, a tab, then the document id or in this case the paragraph id, then a colon, and then the location or the index of the word in the line. The reducer reads the output and then after returns a data structure consisting a dictionary with the word as the index and the value as a dictionary containing the document id as the key and the value as an array with a number for index 0 representing the count of that word in that document and an array at index 1 containing all of the locations or indexes of that word in the given document. An example output line is:

and {'doc2': [6, ['69', '24', '86', '55', '75', '21']], 'doc3': [6, ['71', '66', '11', '51', '15', '27']], 'doc1': [2, ['8', '45']], 'doc4': [5, ['70', '45', '86', '20', '5']], 'doc5': [3, ['80', '31', '57']]}

From that line it can be interpreted that the word and is the 69th, 24th, 86th, 55th, 75th, and 21st word in document 2 and the same can be done for the other documents. From this you are able to tell which word is in which document, the number of times it is seen in that document, and the locations of the word in that document.

The one improvement I would make is to be able to read multiple text files and tag them as different documents rather than each paragraph being considered a new document.