Name:Rhea Adhikari
190905156
Lab5
DS Lab

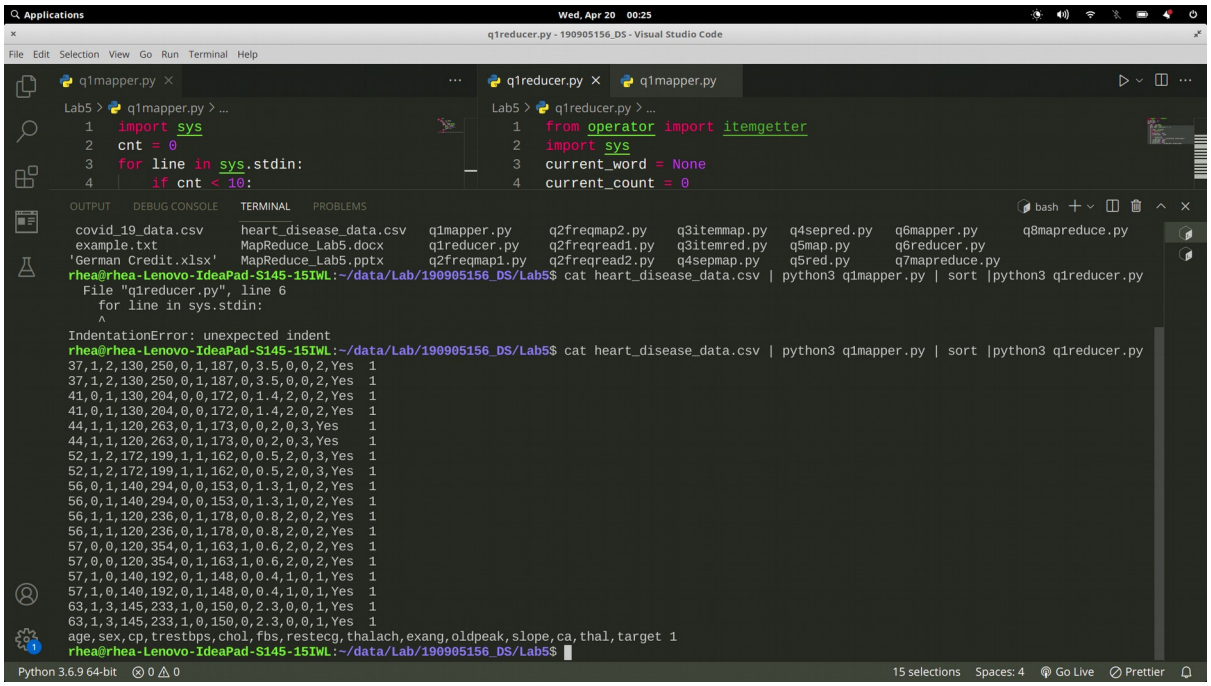
Exercise 1 - Try the above word count program for the Heart Disease dataset, covid_19_data dataset, example dataset and German Credit dataset.

Mapper.py

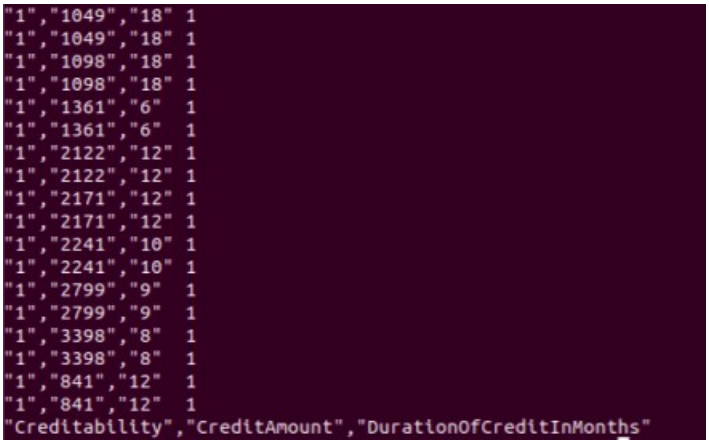
```
import sys
cnt = 0
for line in sys.stdin:
if cnt < 10:
line = line.strip()
words = line.split()
for word in words:
print("%s\t\t%s" %(word, 1))
cnt+=1
else :
break
```

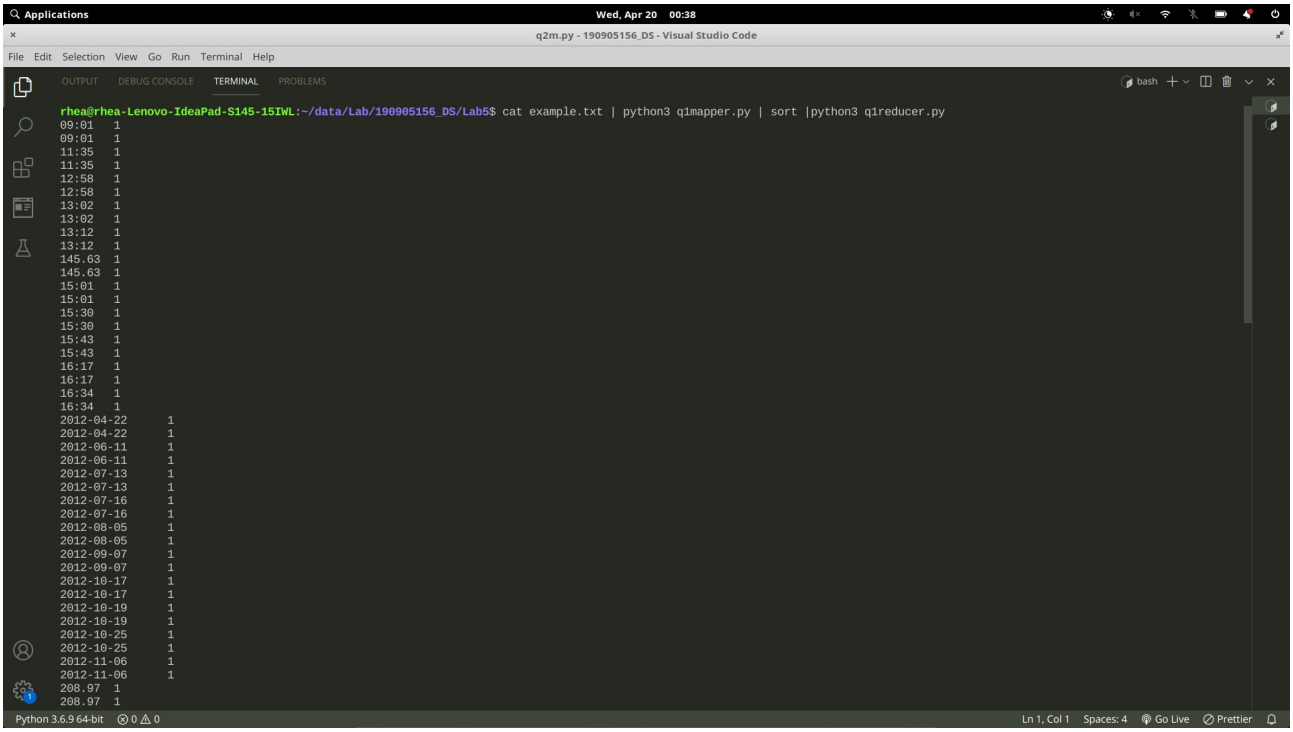
Reducer.py

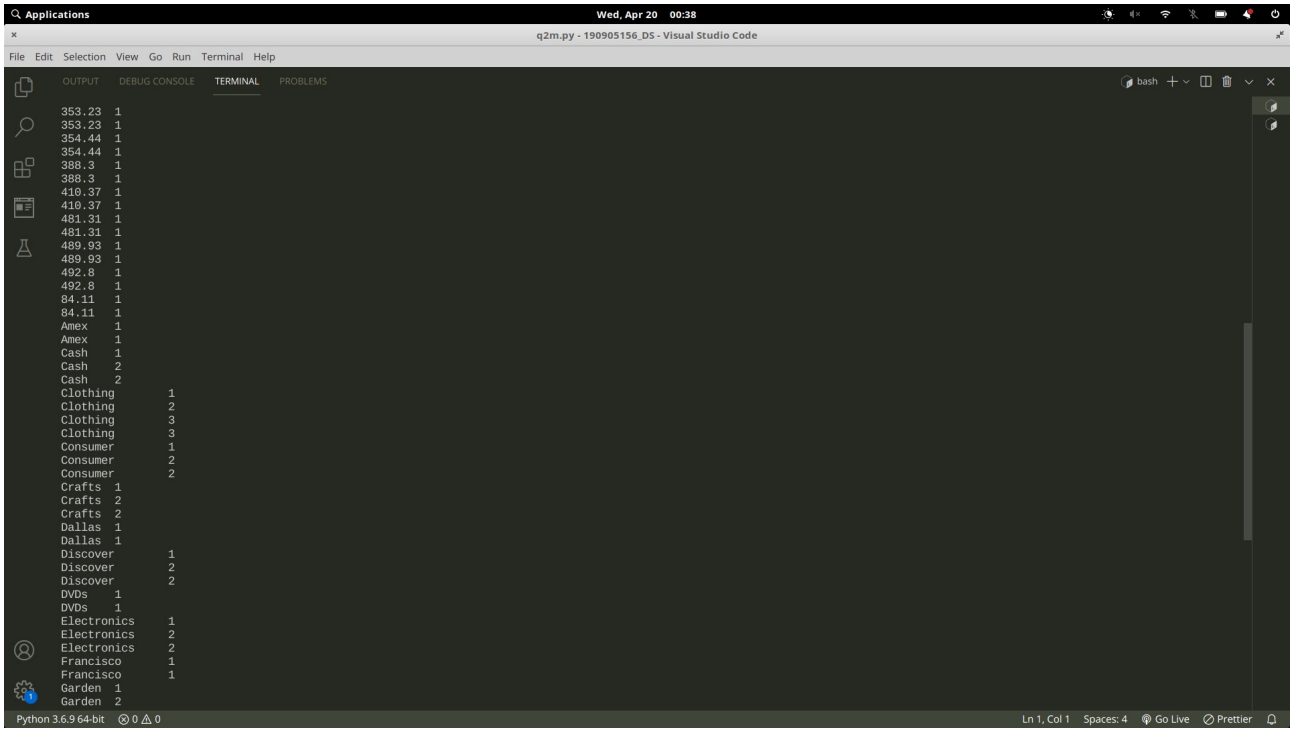
```
from operator import itemgetter
import sys
current_word = None
current_count = 0
word = None
for line in sys.stdin:
line = line.strip()
word, count = line.split('\t', 1)
try:
count = int(count)
except ValueError:
continue
if current_word == word:
current_count += count
else:
if current_word:
print('%s\t%s' % (current_word, current_count) )
current_count = count
current_word = word
if current_word == word:
print('%s\t%s' % (current_word, current_count))
```

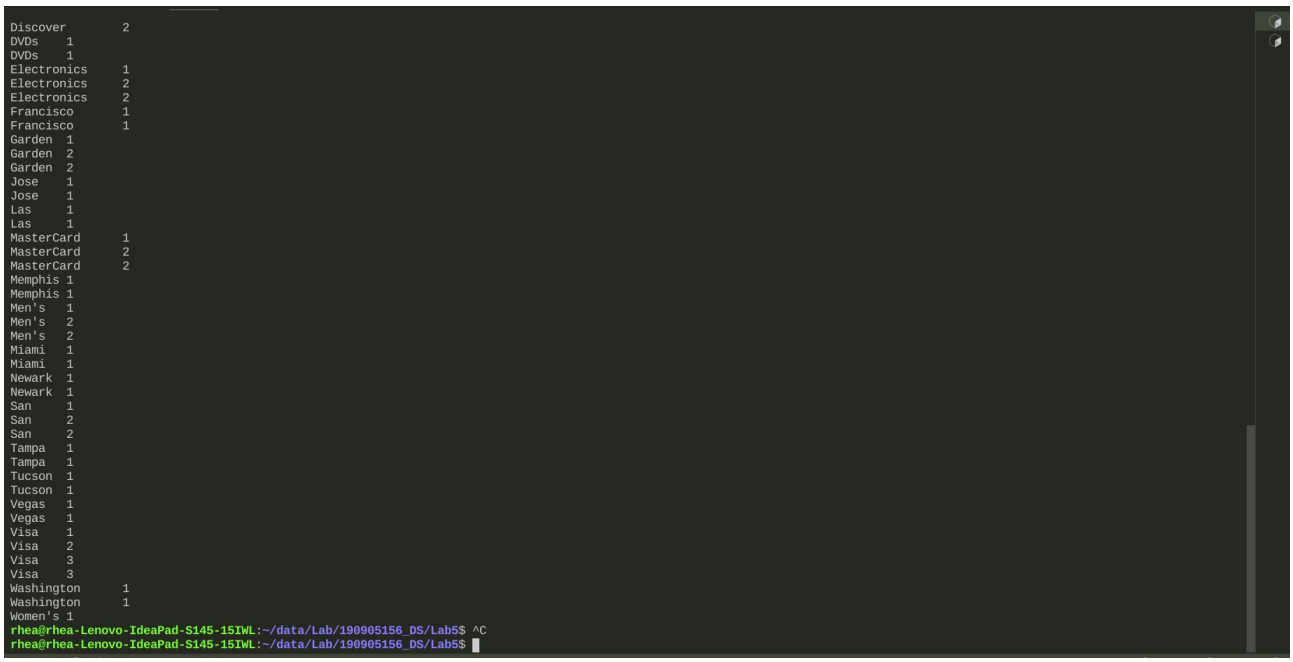


GermanCredit.csv









Exercise 2 : Try the above frequent word count program for the Heart Disease dataset, covid_19_data dataset, example dataset and German Credit data

```
from __future__ import print_function
import sys
for line in sys.stdin:
L = [ (word.strip().lower(), 1 ) for word in line.strip().split() ]
```

```
for word, n in L:
print( '%s\t%d' % (word, n) )


#!/usr/bin/env python
# reducer.py
from __future__ import print_function
import sys
lastWord = None
sum = 0
for line in sys.stdin:
word, count = line.strip().split('\t', 1)
count = int(count)
if lastWord==None:
lastWord = word
sum = count
continue
if word==lastWord:
sum += count
else:
print( "%s\t%d" % ( lastWord, sum ) )
sum = count
lastWord = word
# output last word
if lastWord == word:
print( '%s\t%s' % (lastWord, sum ) )


from __future__ import print_function
import sys
# input comes from STDIN (standard input)
for line in sys.stdin:
word, count = line.strip().split('\t', 1)
count = int(count)
print( '%d\t%s' % (count, word) )

from __future__ import print_function
import sys
mostFreq = []
currentMax = -1
for line in sys.stdin:
count, word = line.strip().split('\t', 1)
count = int(count)
if count > currentMax:
currentMax = count
mostFreq = [ word ]
elif count == currentMax:
mostFreq.append( word )
# output mostFreq word(s)
for word in mostFreq:
print( '%s\t%s' % ( word, currentMax ) )
#combining all the four codes above we get the max frequency elements
```

Applications Wed, Apr 20 00:53

freqred2.py - 190905156_DS - Visual Studio Code

File Edit Selection View Go Run Terminal Help

freqmap1.py freqred1.py freqmap2.py freqred2.py

Lab5 > freqred2.py > ...

```python
    count, word = line.strip().split('\t', 1)
    count = int(count)
    if count > currentMax:
        currentMax = count
        mostFreq = [ word ]
    elif count == currentMax:
        mostFreq.append( word )
    # output mostFreq word(s)
        for word in mostFreq:
            print( '%s\t%s' % ( word, currentMax ) )
```

OUTPUT DEBUG CONSOLE TERMINAL PROBLEMS

```
50,1,2,129,196,0,1,163,0,0,2,0,2,yes    1
50,1,2,140,233,0,1,163,0,0.6,1,1,3,no   1
50,1,2,140,233,0,1,163,0,0.6,1,1,3,no   1
51,0,0,130,305,0,1,142,1,1.2,1,0,3,no   1
51,0,2,120,295,0,0,157,0,0.6,2,0,2,yes  1
51,0,2,120,295,0,0,157,0,0.6,2,0,2,yes  1
51,0,2,130,256,0,0,149,0,0.5,2,0,2,yes  1
51,0,2,130,256,0,0,149,0,0.5,2,0,2,yes  1
51,0,2,140,308,0,0,142,0,1.5,2,1,2,yes  1
51,0,2,140,308,0,0,142,0,1.5,2,1,2,yes  1
51,1,0,140,261,0,0,186,1,0,2,0,2,yes    1
51,1,0,140,261,0,0,186,1,0,2,0,2,yes    1
51,1,0,140,298,0,1,122,1,4.2,1,3,3,no   1
51,1,0,140,298,0,1,122,1,4.2,1,3,3,no   1
51,1,0,140,299,0,1,173,1,1.6,2,0,3,no   1
51,1,0,140,299,0,1,173,1,1.6,2,0,3,no   1
51,1,2,100,222,0,1,143,1,1.2,1,0,2,yes  1
51,1,2,100,222,0,1,143,1,1.2,1,0,2,yes  1
51,1,2,110,175,0,1,123,0,0.6,2,0,2,yes  1
51,1,2,110,175,0,1,123,0,0.6,2,0,2,yes  1
51,1,2,125,245,1,0,166,0,2.4,1,0,2,yes  1
51,1,2,125,245,1,0,166,0,2.4,1,0,2,yes  1
51,1,2,94,227,0,1,154,1,0,2,1,3,yes     1
51,1,2,94,227,0,1,154,1,0,2,1,3,yes     1
51,1,3,125,213,0,0,125,1,1.4,2,1,2,yes  1
51,1,3,125,213,0,0,125,1,1.4,2,1,2,yes  1
52,0,2,136,196,0,0,169,0,0.1,1,0,2,yes  1
52,0,2,136,196,0,0,169,0,0.1,1,0,2,yes  1
```

Python 3.6.9 64-bit ⊗ 0 ⚠ 0 Ln 15, Col 13 Spaces: 4 Go Live Prettier

```
"1","1258","24" 2
"1","1262","12" 2
"1","1374","6"  2
"1","1424","12" 2
"1","1478","15" 2
"1","2171","12" 2
"1","701","12"  2
```

```
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$ cat covid_19_data.csv | python3 freqmap1.py | sort |python3 freqred1.py |python3 freqmap2.py |sort | python3
freqred2.py
and     2675
and     2675
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$
```

3.6.9 64-bit ⊗ 0 ⚠ 0 Ln 16, Col 74 (501 selected) Spaces: 4 Go Live Prettier

```
OUTPUT    DEBUG CONSOLE    TERMINAL    PROBLEMS

09:04   1
09:23   1
09:23   1
09:25   1
09:25   1
09:57   1
09:01   1
09:02   1
09:02   1
09:04   1
09:04   1
09:23   1
09:23   1
09:25   1
09:25   1
09:57   1
09:57   1
amex    10
cash    10
amex    10
cash    10
cash    10
amex    10
cash    10
cash    10
discover        10
amex    10
cash    10
cash    10
discover        10
visa    10
amex    10
cash    10
cash    10
discover        10
visa    10
visa    10
amex    11
discover        11
amex    11
discover        11
discover        11
amex    13
amex    13
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$
```

Exercise 3: Try the above ‘Item explore and count’ for the Heart Disease dataset, covid_19_data dataset, example dataset and German Credit data

```
import fileinput
for line in fileinput.input():
data = line.strip().split("\t")
if len(data) == 6:
date, time, location, item, cost, payment = data
```

```
print ("{0}\t{1}".format(location, cost))
import fileinput
transactions_count = 0
sales_total = 0
for line in fileinput.input():
data = line.strip().split("\t")
if len(data) != 2:
continue
current_key, current_value = data
transactions_count += 1
sales_total += float(current_value)
print (transactions_count, "\t", sales_total)
```

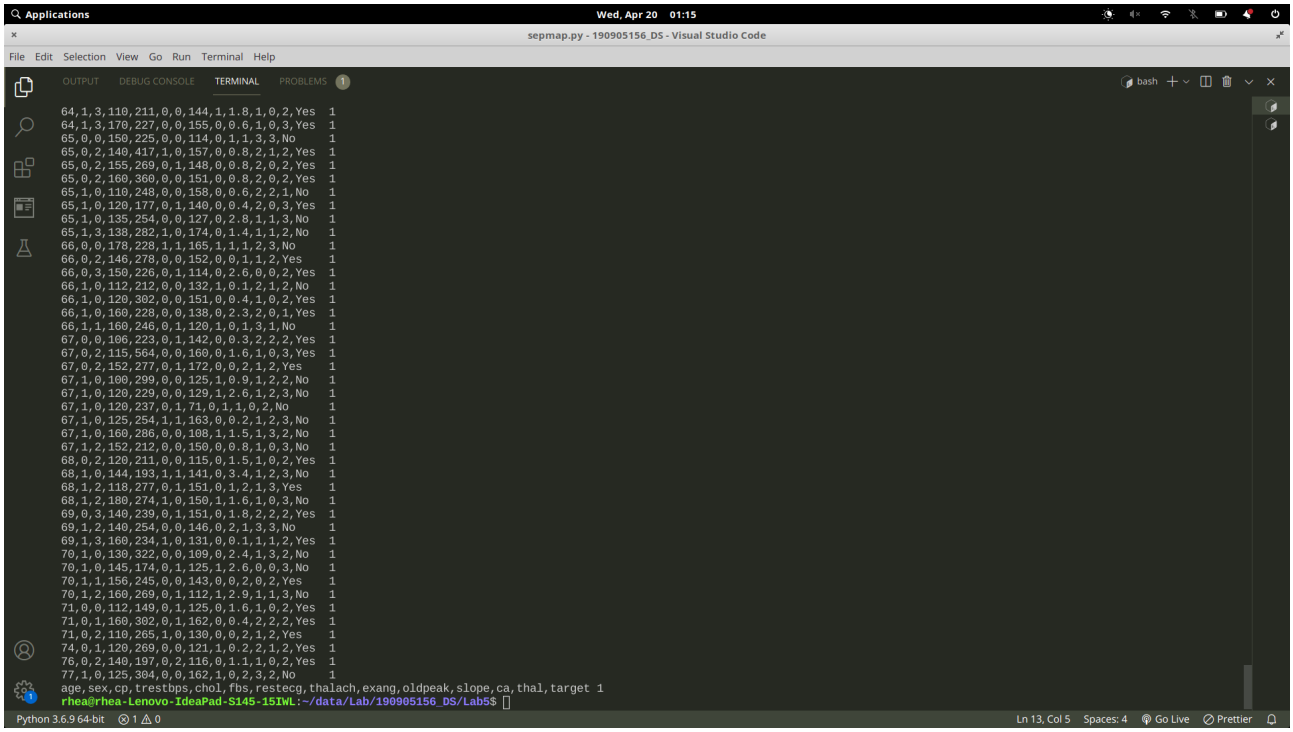
```
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$ cat example.txt | python3 itemmap.py | sort |python3 itemred.py
1        82.38
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$
```

```
290      242.76896551724138
291      242.99656357388315
292      243.2294520547945
293      243.47440273037543
294      243.71768707482994
295      243.96949152542373
296      244.22972972972974
297      244.4915824915825
298      244.76174496644296
299      245.0334448160535
300      245.31666666666666
301      245.61461794019934
302      245.91059602649005
303      246.26402640264027
303      246.26402640264027
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/La
```

Exercise 4: Try to include separator using map reducing for the output of Heart Disease dataset, covid_19_data dataset, example dataset and German Credit data

```
import sys
def read_input(file):
for line in file:
yield line.split()

def main(separator='\t'):
data = read_input(sys.stdin)
for words in data:
for word in words:
print('%s%s%d' % (word, separator, 1))

if __name__=="__main__":
main()
```

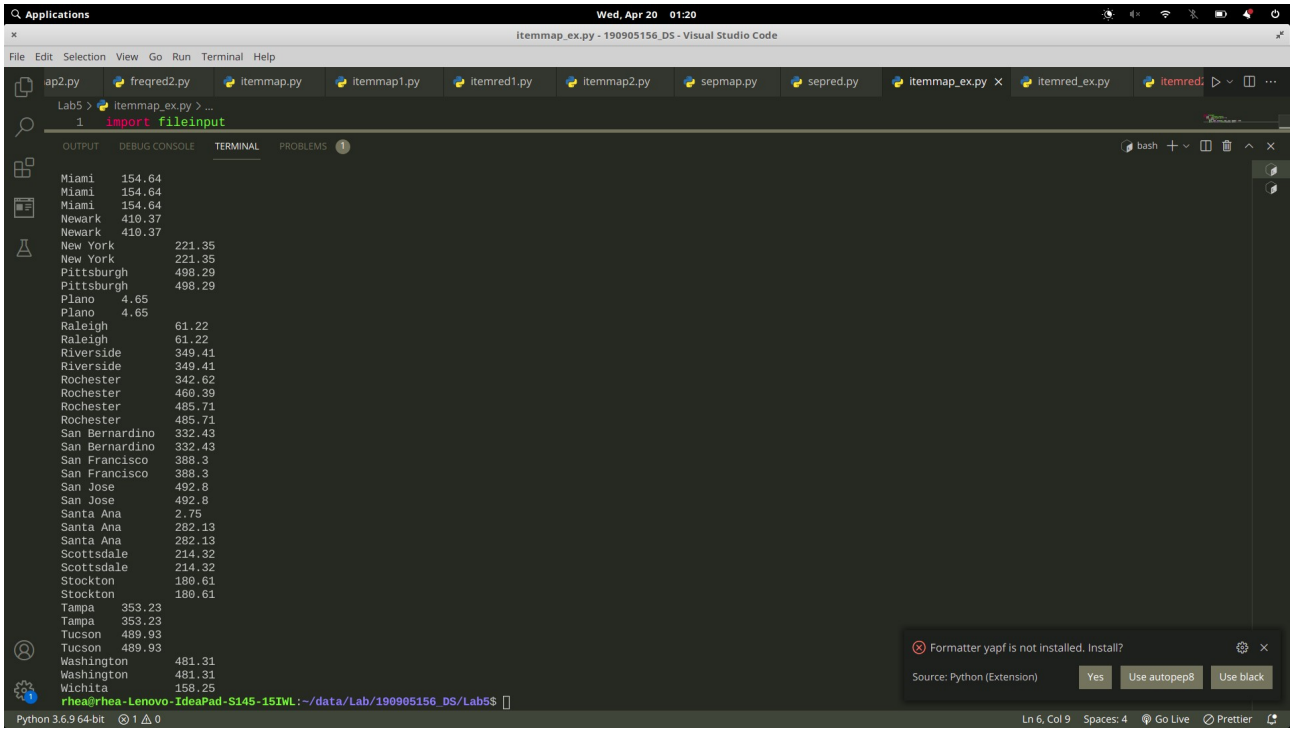
```
from itertools import groupby
from operator import itemgetter
import sys
def read_mapper_output(file, separator='\t'):
for line in file:
yield line.rstrip().split(separator, 1)

def main(separator='\t'):
data = read_mapper_output(sys.stdin, separator=separator)
for current_word, group in groupby(data, itemgetter(0)):
try:
total_count = sum(int(count) for current_word, count in group)
print ("%s%s%d" % (current_word, separator, total_count))
except ValueError:
pass

if __name__=="__main__":
main()
```



Exercise 5: Try to apply finding max value using map reduce concept for the output of Heart Disease dataset, covid_19_data dataset, example dataset and German Credit data

```
import fileinput
for line in fileinput.input():
data = line.strip().split("\t")
if len(data)==6:
date, time, location, item, cost, payment = data
```

```
print("{0}\t{1}".format(location, cost))

import fileinput
max_value = 0
old_key = None
for line in fileinput.input():
data = line.strip().split("\t")
if len(data) != 2:
continue
current_key, current_value = data
if old_key and old_key != current_key:
print(old_key,"\t", max_value)
max_value = 0
if float(current_value) > float(max_value):
max_value = float(current_value)
old_key = current_key
if old_key != None:
print (old_key, "\t", max_value)
```



Exercise 6: TOLD TO SKIP

Exercise 7: Write a map reduce program to count even or odd numbers in randomly generated natural numbers.

```
import fileinput
sum = 0
for line in fileinput.input():
data = line.strip().split("\t")
current_key, current_value = data
sum += int(current_value)
print("Number of odd numbers is:", sum)
```

```
import sys
for line in sys.stdin:
line = line.strip()
words = line.split()
for word in words:
N = int(word)
if N%2 == 1 and N > 0:
print(N, '\t', 1)
```

```
import sys
for line in sys.stdin:
line = line.strip()
words = line.split()
for word in words:
N = int(word)
if N%2 == 0 and N > 0:
print(N, '\t', 1)
```

```
import fileinput
sum = 0
for line in fileinput.input():
data = line.strip().split("\t")
current_key, current_value = data
sum += int(current_value)
print("Number of even numbers is:", sum)
```

```
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/19090515
6_DS/Lab5$ echo "1 2 3 4 5" |python3 oddmap.py|sort|pyt
hon3 oddreduce.py
Number of odd numbers is: 3
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/19090515
6_DS/Lab5$
```

```
6_DS/Lab5$ echo "11 12 321 12 124 21 111" |python3 even
map.py|sort|python3 evenreduce.py
Number of even numbers is: 3
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/19090515
6_DS/Lab5$
```