Anay – 210905071
Lab 05
Solved Exercises

Code
mapper.py

```
import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        print('%s\t%s' % (word, 1))
```

reducer.py

```
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))
```

Sample I/O

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab05-MapReduce-I$ echo
"a a a a v v f f hh hh fg tg fg gt nnn ccc ddd nnn ddd"|python3 Solved_Mapper.py
 | python3 Solved_Reducer.py
a       4
v       2
f       2
hh      2
fg      1
tg      1
fg      1
gt      1
nnn     1
ccc     1
ddd     1
nnn     1
ddd     1
```

FOR heart_disease dataset:
Code
mapper.py

```
import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split(',')
    for word in words:
        print('%s,,%s' % (word, 1))
```

reducer.py

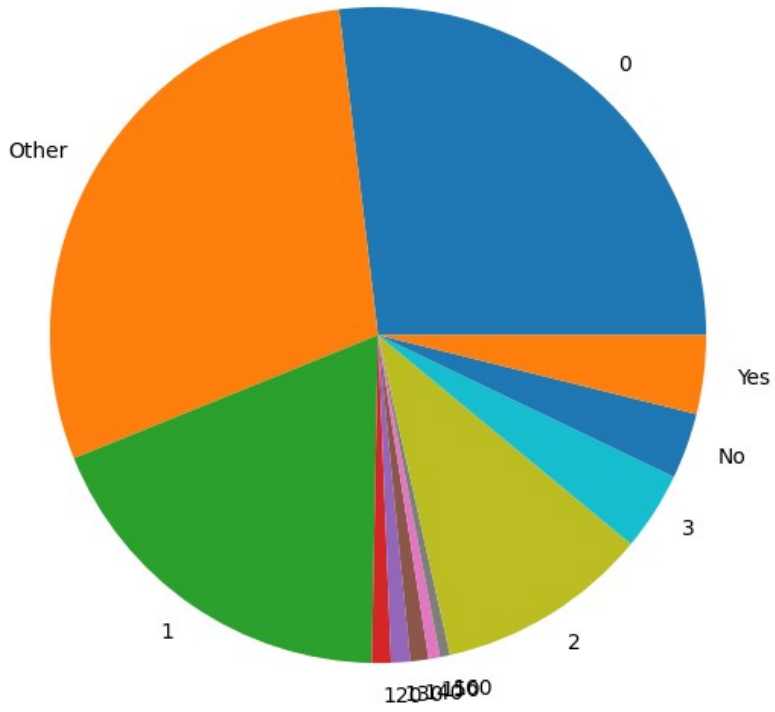
```
import sys

current_word = None
current_count = 0
word = None
for line in sys.stdin:
    line = line.strip()
    word, count = line.split(',,', 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))
```

plotOutput.py

```
from matplotlib import pyplot as plt

in_file = open("out.txt", 'r')
data = {}
for line in in_file:
    inp = line.split('\t')
    if (int(inp[1]) > 20):
        data[inp[0]] = int(inp[1])
    else:
        try:
            if (data["Other"]):
                data["Other"] = data["Other"] + int(inp[1])
        except(Exception):
            data["Other"] = int(inp[1])

fig = plt.figure(figsize=(10, 7))
plt.pie(data.values(), labels=data.keys())
# plt.bar(x=data.keys(), height=data.values())

plt.show()
```

Output

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab05-MapReduce-I$ cat /
home/210905071_Anay/Documents/Distributed\ Systems\ Lab2024/Datasets\ for\ Distr
ibuted\ Systems\ Lab-2024/Lab\ 5\ Required\ Files/heart_disease_data.csv | pytho
n3 Solved_Mapper.py | sort | python3 Solved_Reducer.py > out.txt
```



FOR covid19 dataset:
Code
mapper.py

```
import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split(',')
    for word in words:
        print('%s,,%s' % (word, 1))
```

reducer.py

```
import sys

current_word = None
current_count = 0
word = None
for line in sys.stdin:
    line = line.strip()
    word, count = line.split(',,', 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))
```

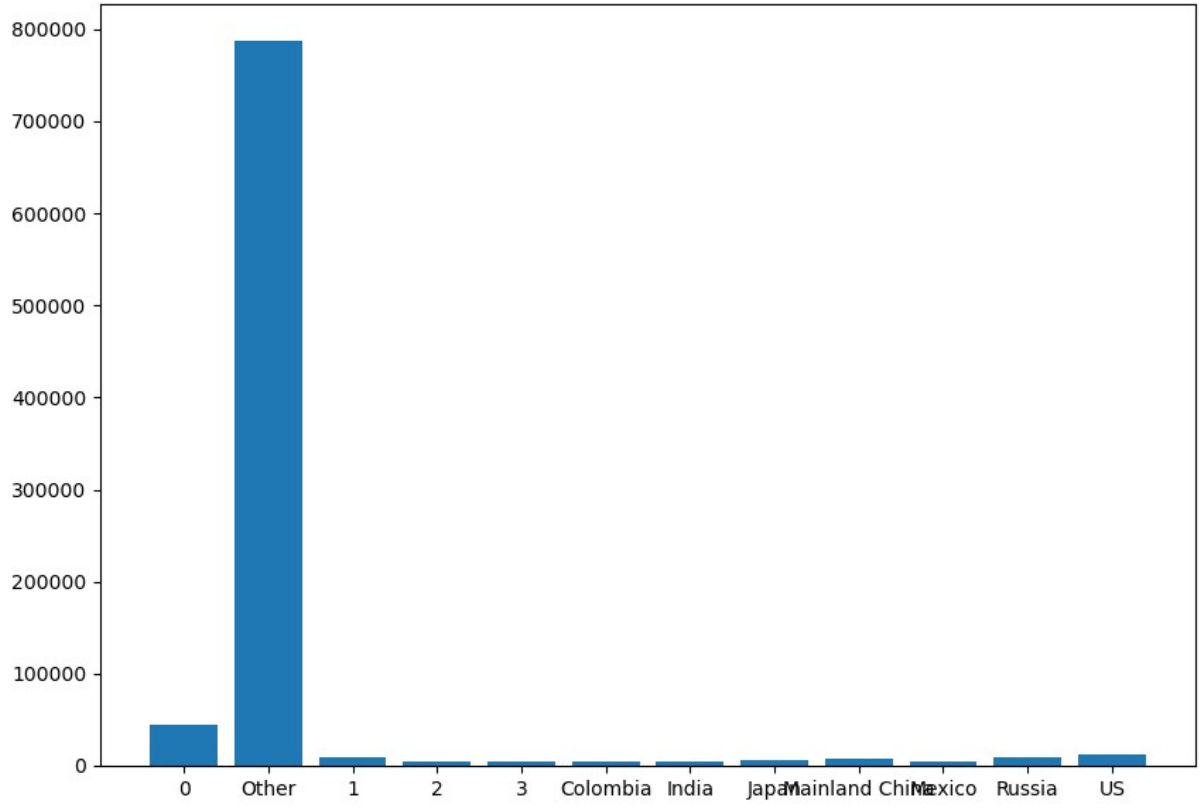
plotOutput.py

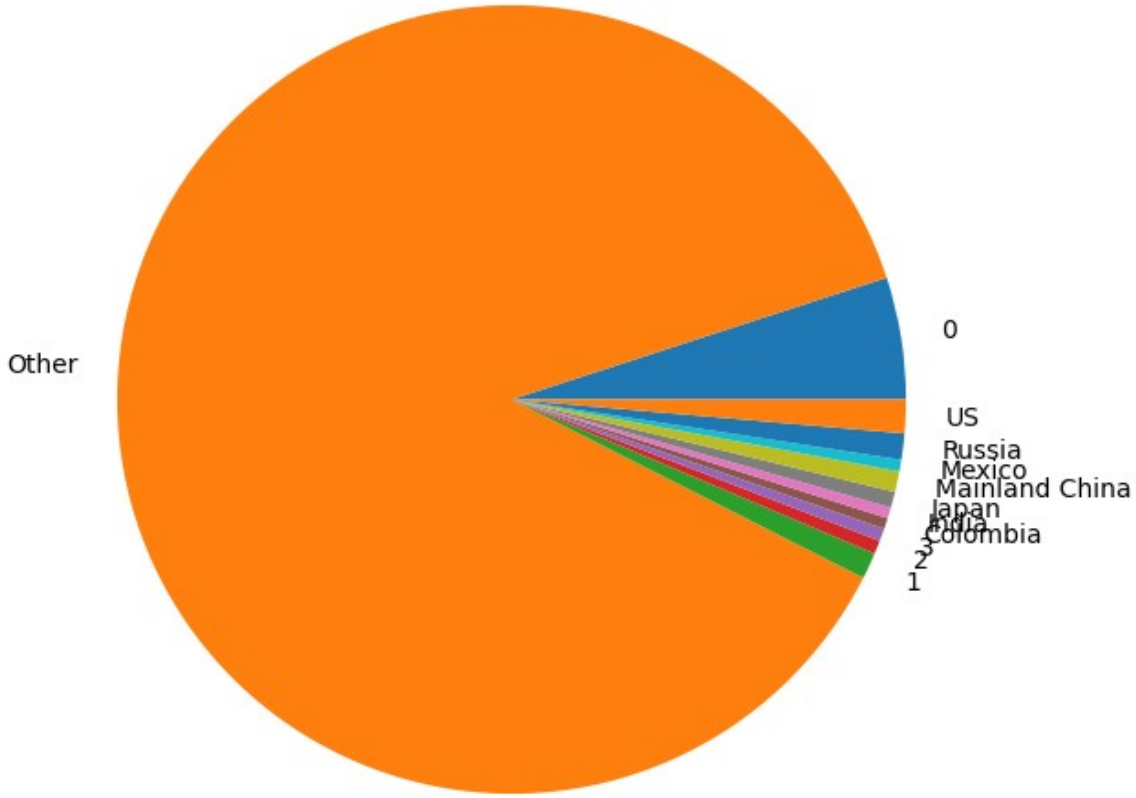
```
from matplotlib import pyplot as plt

in_file = open("out.txt", 'r')
data = {}
for line in in_file:
    inp = line.split('\t')
    if (int(inp[1]) > 4000):
        data[inp[0]] = int(inp[1])
    else:
        try:
            if (data["Other"]):
                data["Other"] = data["Other"] + int(inp[1])
```

```
        except(Exception):
            data["Other"] = int(inp[1])

fig = plt.figure(figsize=(10, 7))
# plt.pie(data.values(), labels=data.keys())
plt.bar(x=data.keys(), height=data.values())

plt.show()
```

Output:

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab05-MapReduce-I$ cat /
home/210905071_Anay/Documents/Distributed\ Systems\ Lab2024/Datasets\ for\ Distr
ibuted\ Systems\ Lab-2024/Lab\ 5\ Required\ Files/covid_19_data.csv | python3 So
lved_Mapper.py | sort | python3 Solved_Reducer.py > out.txt
```





FOR GermanCredit dataset:
Code
mapper.py

```
import sys
import pandas as pd

df = pd.read_excel('GermanCredit.xlsx', engine='openpyxl')

for amount in df["Creditability"]:
    print('%s,,%s' % (amount, 1))
```

reducer.py

```
import sys

current_word = None
current_count = 0
word = None
for line in sys.stdin:
    line = line.strip()
    word, count = line.split(',,', 1)
    try:
```

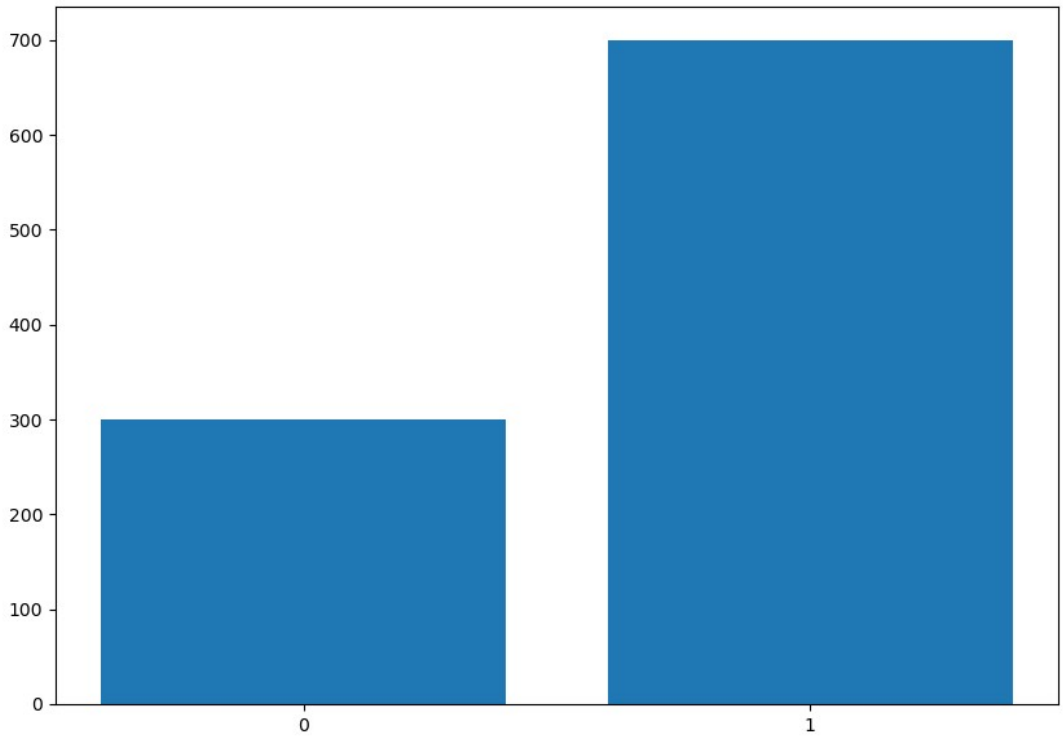
```
        count = int(count)
    except ValueError:
        continue
    if current_word == word:
        current_count += count
    else:
        if current_word:
            print ('%s\t%s' % (current_word, current_count))
        current_count = count
        current_word = word

if current_word == word:
    print ('%s\t%s' % (current_word, current_count))
```

plotOutput.py

```
from matplotlib import pyplot as plt

in_file = open("out.txt", 'r')
data = {}
for line in in_file:
    inp = line.split('\t')
    if (int(inp[1]) > 2):
        data[inp[0]] = int(inp[1])
    else:
        try:
            if (data["Other"]):
                data["Other"] = data["Other"] + int(inp[1])
        except(Exception):
            data["Other"] = int(inp[1])

fig = plt.figure(figsize=(10, 7))
plt.bar(x=data.keys(), height=data.values())

plt.show()
```

Output:

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab05-MapReduce-I$ pytho
n3 Solved_Mapper.py | sort | python3 Solved_Reducer.py > out.txt
```



Frequently used words:
Code
freqmap1.py

```
import sys
for line in sys.stdin:
    line = line.strip()
    L = [ (word.strip().lower(), 1 ) for word in line.strip().split() ]
    for word, n in L:
        print( '%s\t%d' % (word, 1) )
```

freqred1.py

```
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
if lastWord == word:
    print('%s\t%s' % (lastWord, sum ) )
```

freqmap2.py

```
import sys
for line in sys.stdin:
    word, count = line.strip().split('\t', 1)
    count = int(count)
    print( '%d\t%s' % (count, word) )
```

freqred2.py

```
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)

for word in mostFreq:
    print("%s\t%s" % (word, currentMax))
```

Output:

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab05-MapReduce-I$ echo
"foo foo foo labs labs labs quux labs foo bar quux" |python3 freqmap1.py | sort
| python3 freqred1.py
bar     1
foo     4
labs    4
quux    2
```

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab05-MapReduce-I$ echo
"foo foo foo labs labs labs quux labs foo bar quux" |python3 freqmap1.py | sort
| python3 freqred1.py | python3 freqmap2.py
1       bar
4       foo
4       labs
2       quux
```

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab05-MapReduce-I$ echo
"foo foo foo labs labs labs quux labs foo bar quux" |python3 freqmap1.py | sort
| python3 freqred1.py | python3 freqmap2.py | sort
1       bar
2       quux
4       foo
4       labs
```

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab05-MapReduce-I$ echo
"foo foo foo labs labs labs quux labs foo bar quux" |python3 freqmap1.py | sort
| python3 freqred1.py | python3 freqmap2.py | sort | python3 freqred2.py
foo     4
labs    4
```