Lab06 – Anay – 210905071
Solved Exercises

Map reduce with separator

Code:
sepmap.py

```
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()
```

sepred.py

```
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()
```

Output:

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab06-MapReduce-II$ echo
 "Time is gold Time is Time gold" | python3 sepmap.py | sort | python3 sepred.py

gold    2
is      2
Time    3
```

For heart_disease dataset:
Code
sepmap.py

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

def main(separator=',,'):
        data = read_input(sys.stdin)
        for words in data:
            for word in words:
                print ('%s%s%d' % (word, separator, 1))

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

sepred.py

```
from itertools import groupby
from operator import itemgetter
import sys
def read_mapper_output(file, separator=',,'):
    for line in file:
        yield line.rstrip().split(separator, 1)

def main(separator=',,'):
    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, '\t', total_count))
        except ValueError:
            pass

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

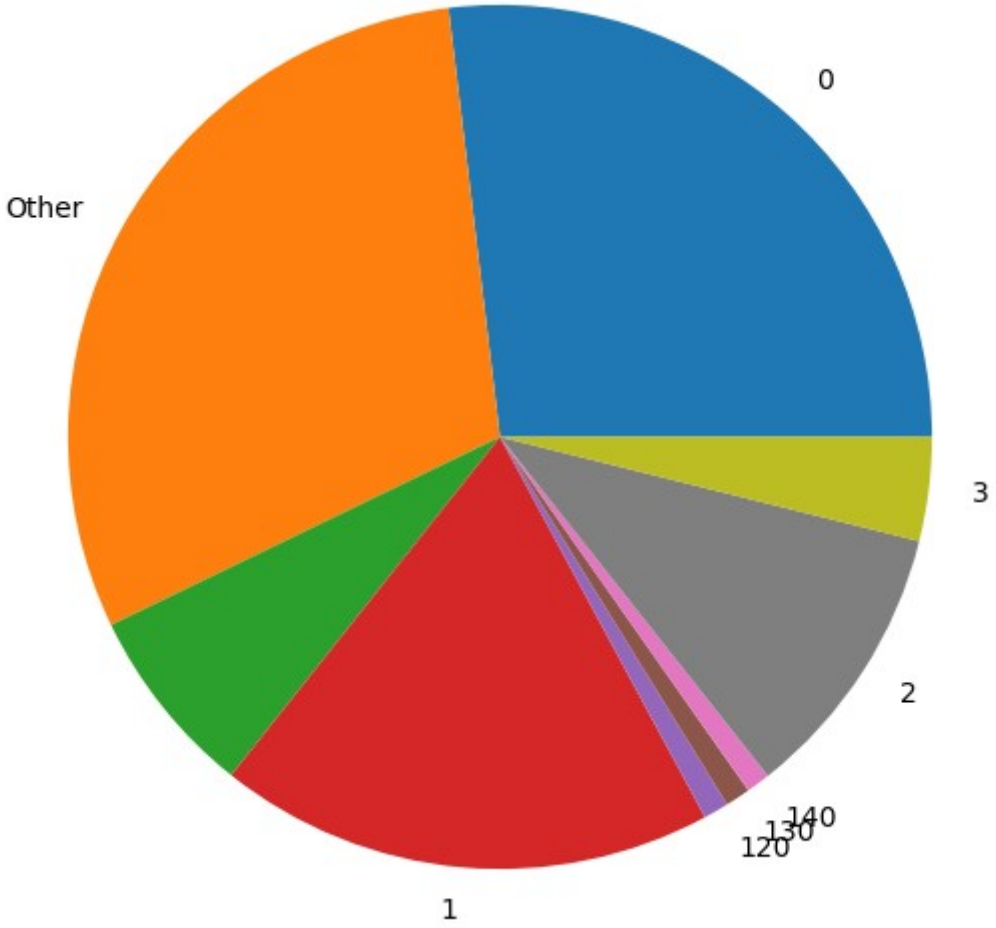
plotOutput.py

```
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]) > 30):
        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/Lab06-MapReduce-II$ cat
/home/210905071_Anay/Documents/Distributed\ Systems\ Lab2024/Datasets\ for\ Dist
ributed\ Systems\ Lab-2024/heart_disease_data.csv | python3 sepmap.py | sort | p
ython3 sepred.py > out.txt
```



For covid_19 dataset for countries in which observations were made:
Code

sepmap.py

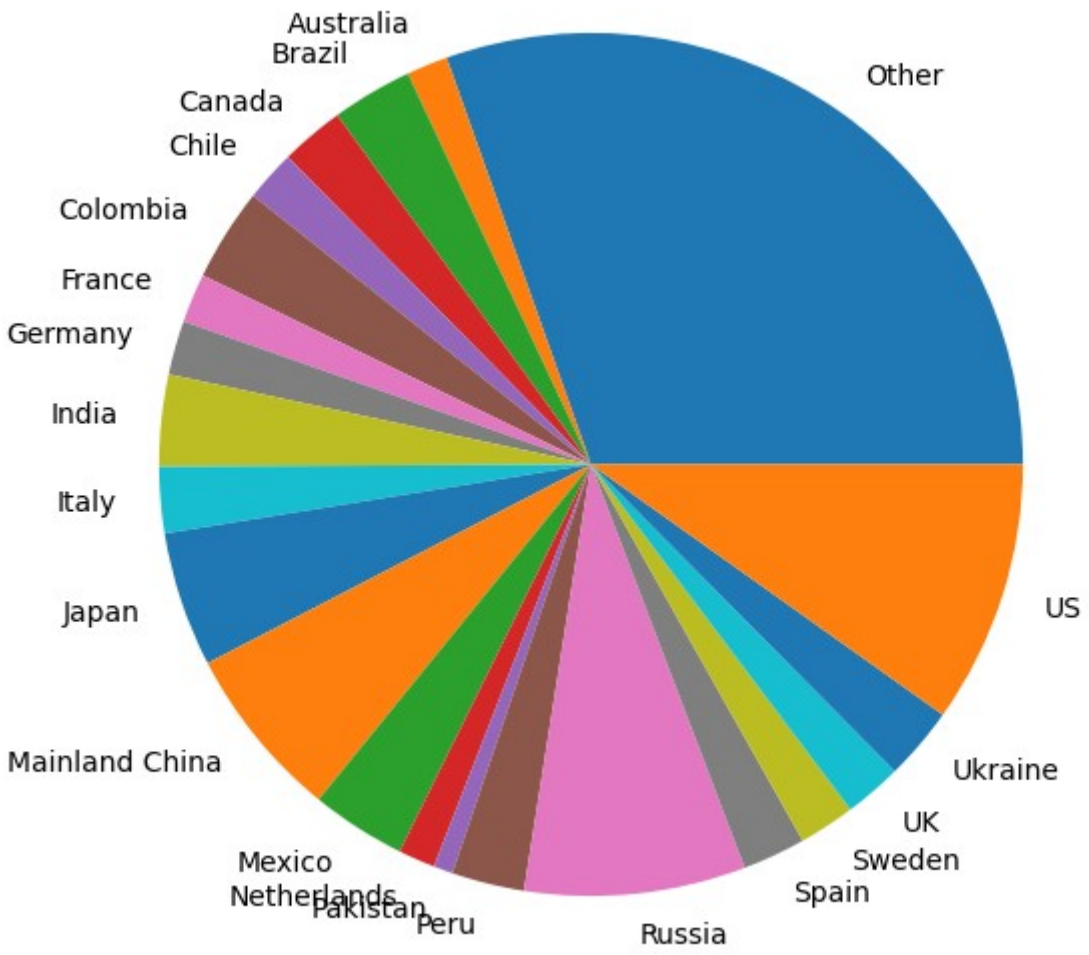
```
import sys
def read_input(file):
    for line in file:
        yield line.split(',')

def main(separator=',,'):
        data = read_input(sys.stdin)
        for words in data:
            word = words[3].strip()
            print ('%s%s%d' % (word, separator, 1))

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

sepred.py

```
from itertools import groupby
from operator import itemgetter
import sys
def read_mapper_output(file, separator=',,'):
    for line in file:
        yield line.rstrip().split(separator, 1)

def main(separator=',,'):
    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, '\t', total_count))
        except ValueError:
            pass

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

Output

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab06-MapReduce-II$ cat
/home/210905071_Anay/Documents/Distributed\ Systems\ Lab2024/Datasets\ for\ Dist
ributed\ Systems\ Lab-2024/covid_19_data.csv | python3 sepmap.py | sort | python
3 sepred.py > out.txt
```



for example dataset:
Code
sepmap.py

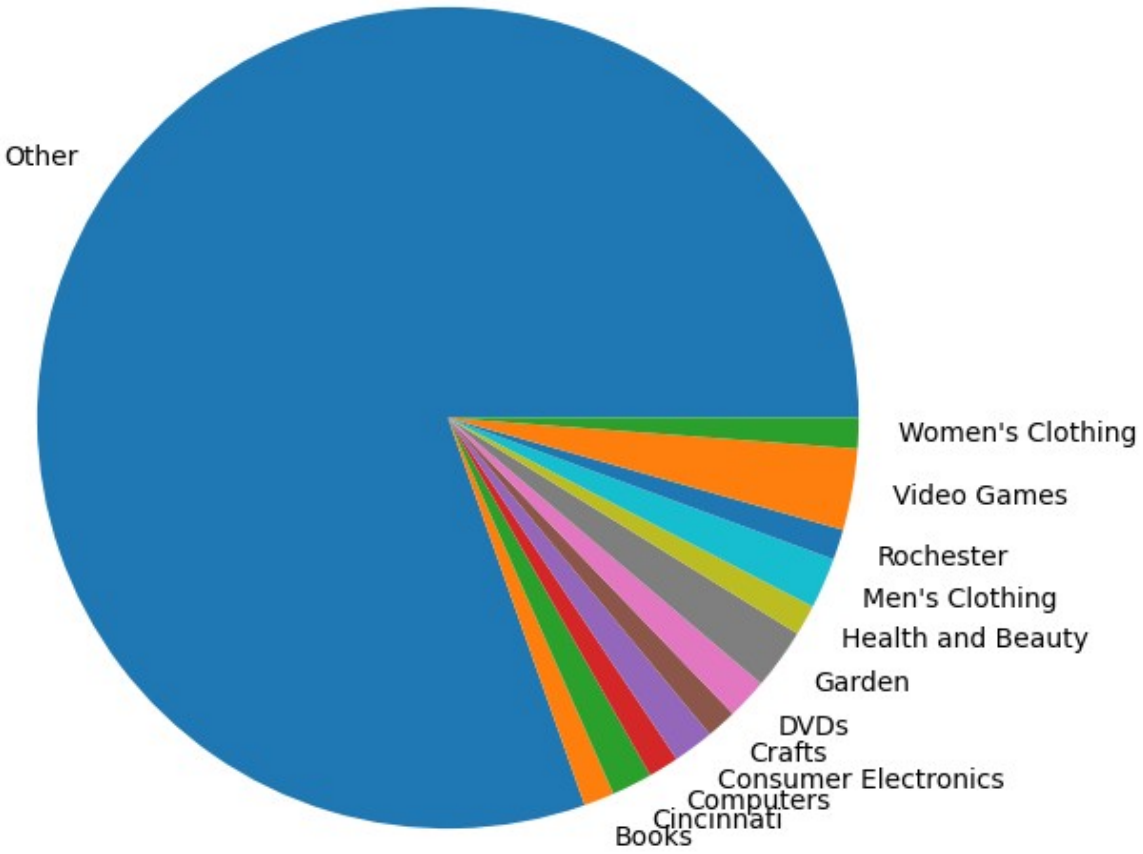
```
import sys
def read_input(file):
    for line in file:
        yield line.split('\t')

def main(separator=',,'):
        data = read_input(sys.stdin)
        for words in data:
            for word in words:
                if (word[-1] == '\n'):
                    continue
                print ('%s%s%d' % (word, separator, 1))

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

sepred.py

```
from itertools import groupby
from operator import itemgetter
import sys
def read_mapper_output(file, separator=',,'):
    for line in file:
        yield line.rstrip().split(separator, 1)

def main(separator=',,'):
    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, '\t', total_count))
        except ValueError:
            pass

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

Output

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab06-MapReduce-II$ cat
/home/210905071_Anay/Documents/Distributed\ Systems\ Lab2024/Datasets\ for\ Dist
ributed\ Systems\ Lab-2024/example.txt | python3 sepmap.py | sort | python3 sepr
ed.py > out.txt
```



For german_credit dataset: (Taking duration of credit values)
Code
sepmap.py

```
import pandas as pd
def main(separator=',,'):
    df = pd.read_excel('GermanCredit.xlsx', engine='openpyxl')

    for amount in df["DurationOfCreditInMonths"]:
        print('%s%s%d' % (amount, separator, 1))

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

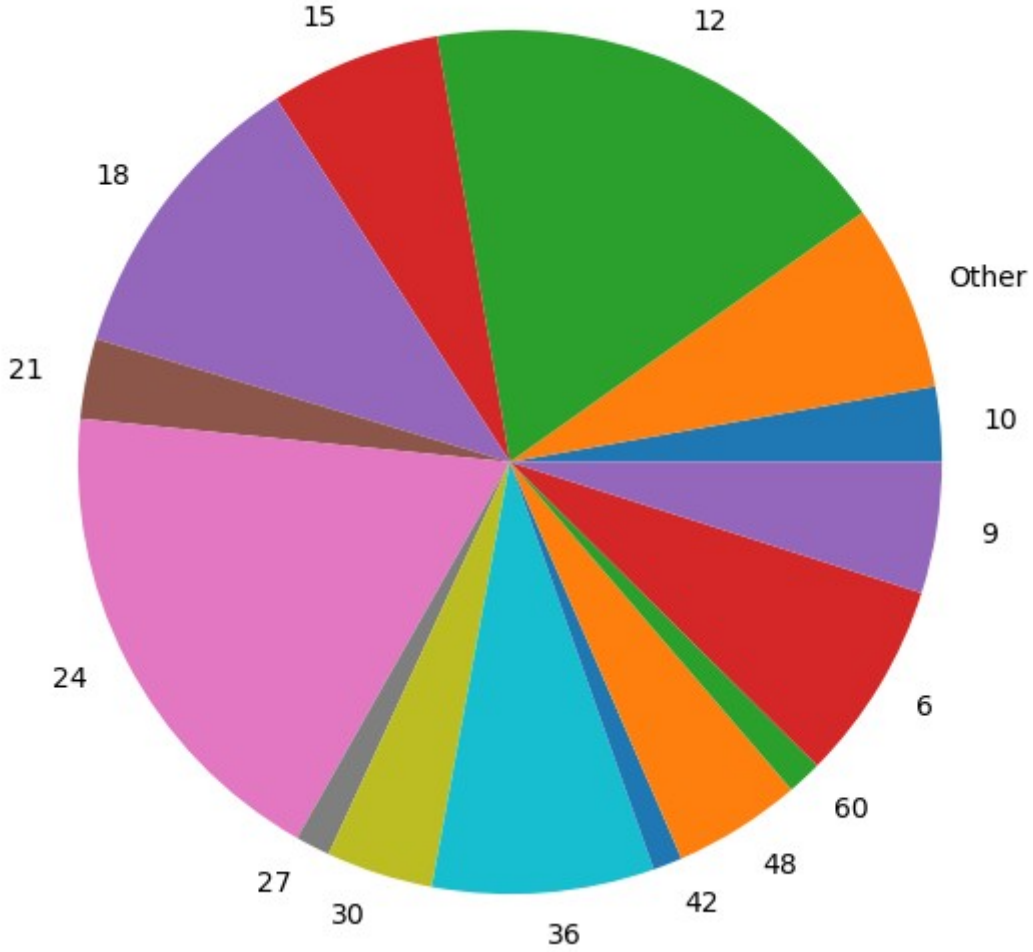
sepred.py

```
from itertools import groupby
from operator import itemgetter
import sys
def read_mapper_output(file, separator=',,'):
    for line in file:
        yield line.rstrip().split(separator, 1)

def main(separator=',,'):
```

```
    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, '\t', total_count))
        except ValueError:
            pass

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

Output

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab06-MapReduce-II$ pyth
on3 sepmap.py | sort | python3 sepred.py > out.txt
```



To find most expensive cost for every city from example.txt dataset
Code

Q5expensive_item_map.py

```
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))
```

Q5expensive_item_red.py

```
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)
```

Output

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab06-MapReduce-II$ cat
/home/210905071_Anay/Documents/Distributed\ Systems\ Lab2024/Datasets\ for\ Dist
ributed\ Systems\ Lab-2024/example.txt | python3 Q5expensive_item_map.py | sort
| python3 Q5expensive_item_red.py
Atlanta          189.22
Aurora   82.38
Austin   48.09
Birmingham       1.64
Boston   397.21
Buffalo          386.56
Chicago          431.73
Cincinnati       443.78
Corpus Christi   157.91
Dallas   145.63
Fremont          404.17
Gilbert          11.31
Glendale         14.09
Indianapolis     464.36
Irvine   15.19
Jersey City      369.07
Las Vegas        208.97
Los      164.5
Louisville       213.64
Lubbock          27.68
Memphis          354.44
Mesa     13.79
Miami    154.64
Newark   410.37
New York         221.35
Pittsburgh       498.29
Plano    4.65
Raleigh          61.22
Riverside        349.41
Rochester        485.71
San Bernardino   332.43
San Francisco    388.3
San Jose         492.8
Santa Ana        282.13
```

For heart_disease dataset:
Code
Q5map.py

```
import fileinput
for line in fileinput.input():
    data = line.strip().split(",")
    if len(data) == 14:
        age, sex, cp, trestbps, chol, fbs, restecg, thalach,exang, oldpeak, slope, ca, thal, target = data
        print ("{0}\t{1}".format(age, chol))
```

Q5red.py

```
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)
```

Output

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab06-MapReduce-II$ cat
/home/210905071_Anay/Documents/Distributed\ Systems\ Lab2024/Datasets\ for\ Dist
ributed\ Systems\ Lab-2024/heart_disease_data.csv | python3 Q5map.py | sort | py
thon3 Q5red.py
29       204.0
34       210.0
35       282.0
37       250.0
38       231.0
39       321.0
40       223.0
41       306.0
42       315.0
43       341.0
44       290.0
45       309.0
46       311.0
47       275.0
48       275.0
49       271.0
50       254.0
51       308.0
52       325.0
53       282.0
54       309.0
55       353.0
56       409.0
57       354.0
58       340.0
59       326.0
60       318.0
61       330.0
62       394.0
63       407.0
64       335.0
65       417.0
66       302.0
67       564.0
68       277.0
69       254.0
70       322.0
```

Q6. Get max confirmed number of cases for each country:
Code
Q6map.py

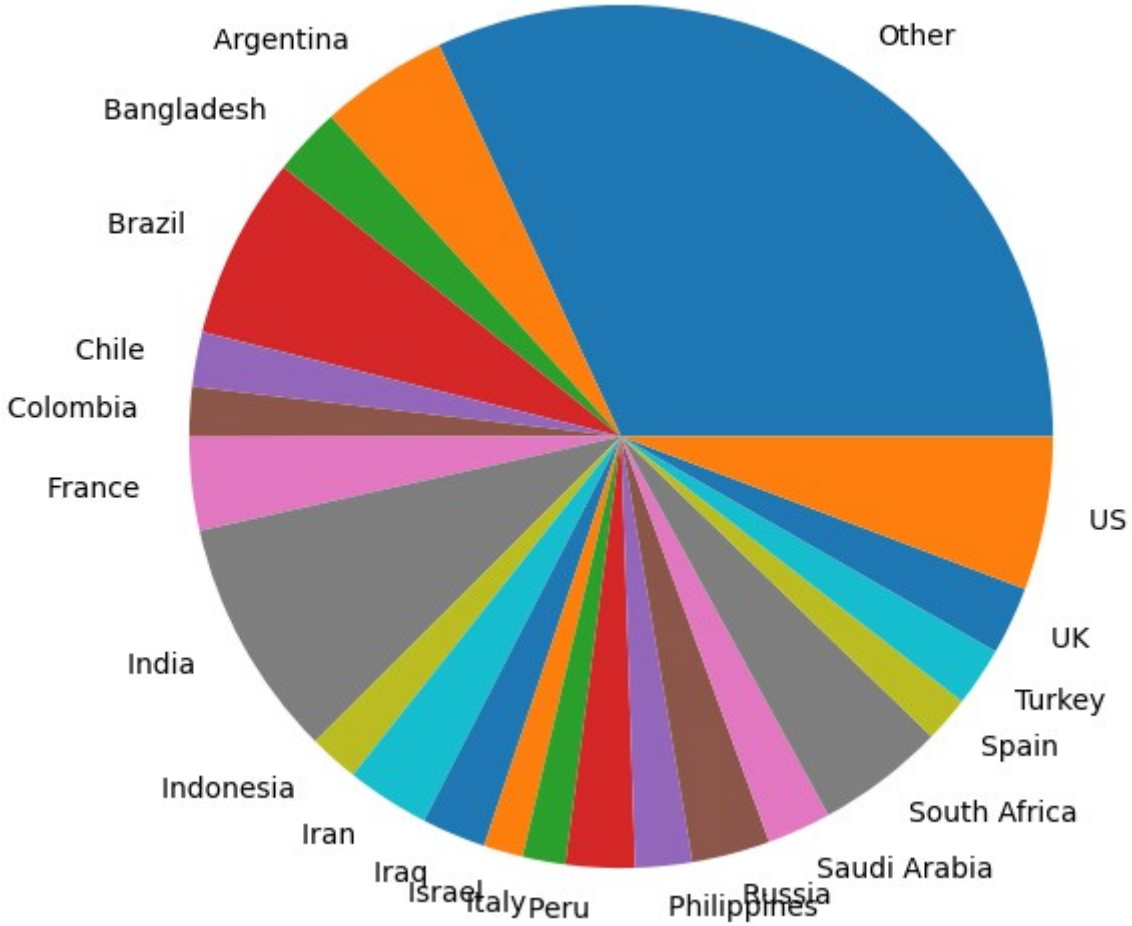
```
import fileinput
ind = 0
for line in fileinput.input():
    data = line.strip().split(",")
    if ind == 0:
        ind += 1
        continue
    if len(data)==8:
        _, _, _, country, _, confirmed_count, _, _ = data
        print("{0}\t{1}".format(country, confirmed_count))
```

Q6red.py

```
import fileinput
max_value = -1
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 = -1
    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)
```

Output

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab06-MapReduce-II$ cat
/home/210905071_Anay/Documents/Distributed\ Systems\ Lab2024/Datasets\ for\ Dist
ributed\ Systems\ Lab-2024/covid_19_data.csv | python3 Q6map.py | sort | python3
 Q6red.py > out.txt
```



Q7. Program to count number of even and odd numbers in randomly generated numbers:
Code
Q7map.py

```
import random

with open('random_numbers.txt', 'w') as file:
    for i in range(500):
        n = random.randint(1, 10000)
        file.write(str(n) + "\n")

with open('random_numbers.txt', 'r') as file:
    for line in file:
        num = line.strip()[:-1] # to remove the \n
        if (num == ''):
            continue
        if (int(num) % 2 == 0):
            print(f"Even\t{1}")
        else:
            print(f"Odd\t{1}")
```

Q7red.py

```
import sys

file = sys.stdin

dict = {}

for line in file:
    data = line.split('\t')
    if data[0] in dict.keys():
        dict[data[0]] += 1
    else:
        dict[data[0]] = 0

for key in dict.keys():
    print(f"{key}\t{dict[key]}")
```

Output:

```
210905071_Anay@netwotklab:~/Documents/CSESem6Labs/DSLab/Lab06-MapReduce-II$ pyth
on3 Q7map.py | sort | python3 Q7red.py
Even    223
Odd     275
```