Student Name : Singh Jasraj
Group : A29
Date : 21 March, 2023

LAB 4: ANALZING NETWORK DATA LOG

You are provided with the data file, in .csv format, in the working directory. Write the program to extract the following informations.

EXERCISE 4A: TOP TALKERS AND LISTENERS

One of the most commonly used function in analyzing data log is finding out the IP address of the hosts that send out large amount of packet and hosts that receive large number of packets, usually know as TOP TALKERS and LISTENERS. Based on the IP address we can obtained the organization who owns the IP address.

TOP 5 TALKERS

Rank	IP address	# of packets	Organisation
1	193.62.192.8	3041	European Bioinformatics Institute
2	155.69.160.32	2975	Nanyang Technological University
3	130.14.250.11	2604	National Library of Medicine
4	14.139.196.58	2452	Indian Institute of Technology
5	140.112.8.139	2056	Taiwan Academic Network / National Taiwan University

TOP 5 LISTENERS

Rank	IP address	# of packets	Organisation
1	103.37.198.100	3841	A*STAR
2	137.132.228.15	3715	National University of Singapore
3	202.21.159.244	2446	Rpnet
4	192.101.107.153	2368	Battelle Memorial Institute, Pacific Northwest Division
5	103.21.126.2	2056	Indian Institute of Technology, Bombay

EXERCISE 4B: TRANSPORT PROTOCOL

Using the IP protocol type attribute, determine the percentage of TCP and UDP protocol

	Header value	Transport layer protocol	# of packets
1	6	TCP	56064 (80.82%)
2	17	UDP	9462 (13.64%)

EXERCISE 4C: APPLICATIONS PROTOCOL

Using the Destination IP port number determine the most frequently used application protocol. (For finding the service given the port number https://www.adminsub.net/tcp-udp-port-finder/)

Rank	Destination IP port number	# of packets	Service
1	443	13423	HTTPS/SSL
2	80	2647	HTTP
3	52866	2068	Dynamic Port
4	45512	1356	Unassigned
5	56152	1341	Dynamic Port

EXERCISE 4D: TRAFFIC

The traffic intensity is an important parameter that a network engineer needs to monitor closely to determine if there is congestion. You would use the IP packet size to calculate the estimated total traffic over the monitored period of 15 seconds. (Assume the sampling rate is 1 in 2048)

Total Traffic (MB)	126519.184 MB

EXERCISE 4E: ADDITIONAL ANALYSIS

Please append ONE page to provide additional analysis of the data and the insight it provides.

TOP 5 COMMUNICATION PAIRS

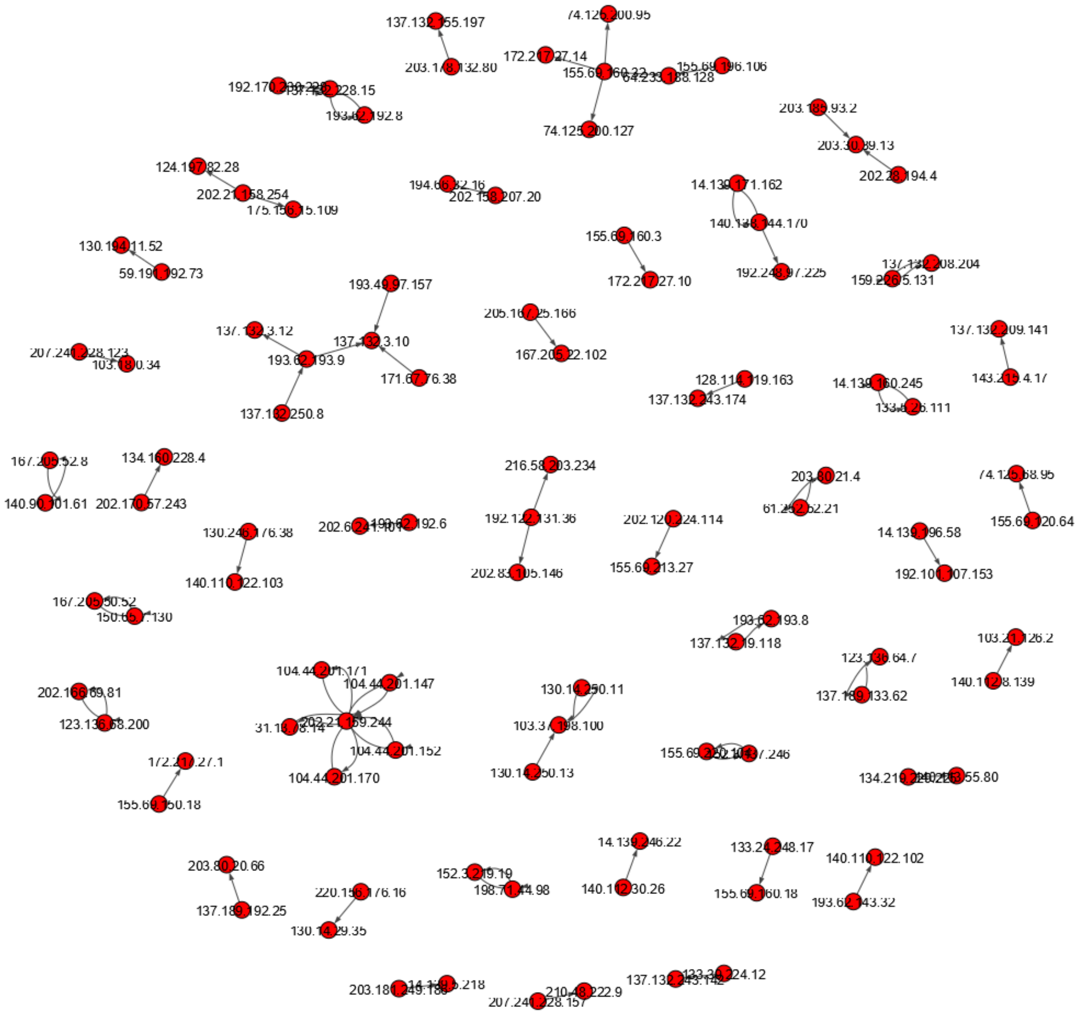
Rank	Source IP	Source Org	Destination IP	Destination Org	Count
1	193.62.192.8	European Bioinformatics Institute	137.132.228.15	National University of Singapore	3041
2	130.14.250.11	National Library of Medicine	103.37.198.100	A*STAR	2599
3	14.139.196.58	Indian Institute of Technology	192.101.107.153	Battelle Memorial Institute, Pacific Northwest Division	2368
4	140.112.8.139	National Taiwan University	103.21.126.2	Indian Institute of Technology Bombay	2056
5	137.132.228.15	National University of Singapore	193.62.192.8	European Bioinformatics Institute	1910

VISUALIZATION OF TOP 80 COMMUNICATION PAIRS

See next page

EXERCISE 4F: SOFTWARE CODE

Please also submit your code to the NTULearn lab site.


137.132.155.197
203.178.132.80
74.125.200.95
172.217.27.14
155.69.160.22
64.233.188.128
155.69.196.106
74.125.200.127
192.170.230.228
137.132.228.15
193.62.192.8
203.185.93.2
203.30.39.13
202.28.194.4
124.197.82.28
202.21.158.254
175.156.15.109
194.66.82.16
202.158.207.20
14.139.171.162
140.138.144.170
155.69.160.3
172.217.27.10
192.248.97.225
130.194.11.52
59.191.192.73
137.132.208.204
159.226.5.131
193.49.97.157
205.167.25.166
137.132.3.12
137.132.3.10
167.205.22.102
193.62.193.9
207.241.228.123
103.180.0.34
171.67.76.38
137.132.209.141
128.114.119.163
137.132.243.174
143.215.4.17
14.139.160.245
133.6.26.111
137.132.250.8
167.205.52.8
134.160.228.4
216.58.203.234
203.80.21.4
74.125.68.95
140.90.101.61
202.170.57.243
61.252.62.21
192.122.131.36
202.120.224.114
155.69.120.64
202.6.241.101
193.62.192.6
130.246.176.38
14.139.196.58
140.110.122.103
202.83.105.146
155.69.213.27
192.101.107.153
167.205.50.52
150.65.7.130
193.62.193.8
137.132.19.118
103.21.126.2
123.136.64.7
104.44.201.171
104.44.201.147
140.112.8.139
130.14.250.11
137.189.133.62
202.166.69.81
31.13.78.14
202.21.159.244
123.136.68.200
103.37.198.100
104.44.201.152
155.69.252.104
202.1.37.246
134.219.240.202
202.158.55.80
172.217.27.1
104.44.201.170
130.14.250.13
155.69.150.18
14.139.246.22
133.24.248.17
140.110.122.102
203.80.20.66
152.3.219.19
220.156.176.16
198.71.44.98
140.112.30.26
155.69.160.18
137.189.192.25
193.62.143.32
130.14.29.35
14.139.5.218
137.132.243.142
133.39.224.12
203.181.249.188
210.48.222.9
207.241.228.157

Setup

```
import pandas as pd
import requests

from igraph import Graph, plot
```

```
cols = [
    "type", "sflow_agent_address", "input_port", "output_port", "src_mac", \
    "dst_mac", "ethernet_type", "in_vlan", "out_vlan", "src_ip", \
    "dst_ip", "ip_protocol", "ip_tos", "ip_ttl", "src_port", \
    "dst_port", "tcp_flags", "packet_size", "ip_size", "sampling_rate", \
]
df = pd.read_csv("./Lab4-ActualData.csv", usecols=range(20), header=None, names=cols)

df.head()
```

```
def get_org(ip):
    return requests.get(f"http://ip-api.com/json/{ip}").json()["org"]
```

EXERCISE 4A: TOP TALKERS AND LISTENERS

TOP TALKERS

```
talkers = df.loc[:, ["src_ip"]] \
    .value_counts() \
    .nlargest(5) \
    .reset_index() \
    .rename({0: "count"}, axis=1)

talkers.loc[:, "src_org"] = talkers.loc[:, "src_ip"].apply(get_org)

talkers = talkers.loc[:, ["src_ip", "src_org", "count"]]

talkers
```

TOP LISTENERS

```
listeners = df.loc[:, ["dst_ip"]] \
    .value_counts() \
    .nlargest(5) \
    .reset_index() \
    .rename({0: "count"}, axis=1)

listeners.loc[:, "dst_org"] = listeners.loc[:, "dst_ip"] \
    .apply(get_org)

listeners = listeners.loc[:, ["dst_ip", "dst_org", "count"]]

listeners
```

EXERCISE 4B: TRANSPORT PROTOCOL

```
ip_type = df.loc[:, ["ip_protocol"]] \
    .value_counts()[[6, 17]] \
    .reset_index() \
    .rename({0: "count"}, axis=1)

ip_type.loc[:, "percentage"] = 100*ip_type.loc[:, "count"] / df.shape[0]

ip_type
```

EXERCISE 4C: APPLICATIONS PROTOCOL

```
df[["dst_port"]] \
    .value_counts() \
    .nlargest(5) \
    .reset_index() \
    .rename({0: "count"}, axis=1)
```

EXERCISE 4D: TRAFFIC

```
size_recorded = df["ip_size"].sum() / (1024**2)
total_size = size_recorded * 2048

print(f"Total traffic: {total_size:.3f} MB")
```

EXERCISE 4E: ADDITIONAL ANALYSIS

TOP 5 COMMUNICATION PAIRS

```
top_comm_pairs = df.groupby(["src_ip", "dst_ip"]) \
    .size() \
    .sort_values(ascending=False) \
    .nlargest(5) \
    .reset_index() \
    .rename({0: "count"}, axis=1)

top_comm_pairs.loc[:, "src_org"] = top_comm_pairs.loc[:, "src_ip"].apply(get_org)
top_comm_pairs.loc[:, "dst_org"] = top_comm_pairs.loc[:, "dst_ip"].apply(get_org)

top_comm_pairs = top_comm_pairs.loc[:, ["src_ip", "src_org", "dst_ip", "dst_org", "count"]]
top_comm_pairs.to_csv("top_comm_pairs.csv")

top_comm_pairs
```

GRAPHICAL VISUALISATION OF TOP 80 COMMUNICATING PAIRS

```
comm_pairs = df.groupby(["src_ip", "dst_ip"]) \
    .size() \
    .nlargest(80)

comm_pairs.head()
```

```
edges = [tuple(edge) for edge in comm_pairs.index]
graph = Graph.TupleList(edges, directed=True, weights=False)

graph.vs["size"] = 15
graph.vs["label_size"] = 12
graph.vs["label"] = graph.vs["name"]
graph.es["arrow_size"] = 0.5

plot(graph, "network_vis.png", margin=50, bbox=(1000, 1000))
```