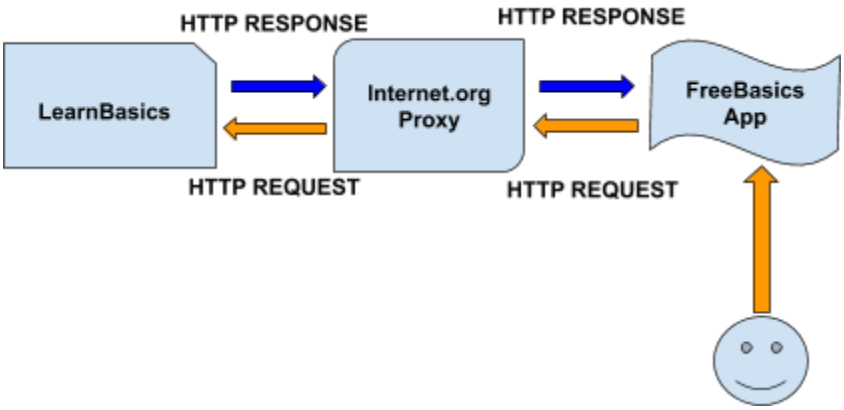
CS 39006: Networks Lab
Assignment 2:
Packet Sniffer and Packet Analyzer -- Exploring Further
Learning `tshark` and writing scripts to analyze pcap
Submission Deadline: January 22, 2020, 2:00 PM

I guess all of you the story behind the Internet.org and FreeBasics services (https://en.wikipedia.org/wiki/Internet.org) that was developed and popularized by Facebook and have seen a lot of debates later on before being banned in many countries in India. In this assignment, we'll analyze a traffic trace collected while using the Freebasics service to explore a few of the interesting stuff regarding the range and scope of the Freebasics service.

One of the important aspects of the FreeBasics service is that how many users from which countries have accessed the service. For this purpose, we had developed a web service called LearnBasics (a simple website for teaching basic English words to students), deployed it on a cloud server, and registered the service with FreeBasics. The domain name for this service was http://learnbasics.mpi-sws.org (the domain is not active right now). The users having the FreeBasics app installed on their mobile can access this LearnBasics service through Facebook's FreeBasics app.

FreeBasics works in the following way (see the following diagram). Facebook used to run a FreeBasics HTTP proxy (Internet.org) in different parts of the globe. When a user wants to access a FreeBasics service (like LearnBasics), the request goes through the FreeBasics Proxy and the ISP makes that service free for the user.



We have collected a pcap trace at the LearnBasics server. The pcap trace is available here -- https://drive.google.com/drive/folders/1VFTwFDkFAsACXJfyBGCUtr7-XexWiGnG?usp=sharing

Your task will be to write a python script to find out the answer to the following questions.

(a) What are the different countries from which the LearnBasics service has been accessed via FreeBasics?
(b) How many users from each of those countries have accessed the service via FreeBasics?

Execute the following steps to do this.

1. Use the `tshark` tool to extract only the HTTP Requests from the the pcap file and generate a new pacp with the filtered HTTP Request packets.
2. Use the `tshark` tool to covert the new pcap (with only HTTP Request packets) to a xml file.
3. Write a python script that will take this xml file as the input and will produce a csv file in the output. The csv file will have the following two fields --
    a. Country name
    b. Number of users who accessed the service from this country

Sample Input and Output:

Say, the xml file name is `http.xml`, and the python script is parse.py.

```
>> python parse.py http.xml
The output is written in data.csv
```

The file `data.csv` will contain something as follows.

KENYA 112
COLUMBIA 32
SENEGAL 34
EGYPT 99
IRAQ 56

Important Points:

1. Note that the request has been forwarded by a HTTP proxy, and the trace has been collected at the server. So, the Source IP field at the IP header may not contain the IP of the user. You should check the trace and find out which field contains the IP of the user.

2. Some users might directly access the service via the web, and not via the FreeBasics app (not through Internet.org proxy). We are only interested in the users who accessed the service via the Internet.org proxy. You should check this.

3. You can use the `python-geoip` package to find out the geolocation for an IP, which returns the country code.

4. You can use the `pycountry` package to find out the country name for a country code.

Submission Instruction:

You need to submit the python source via Moodle by the deadline. Your Python source should have sufficient comments explaining the steps you have used in the code.