Preprocess Dataset

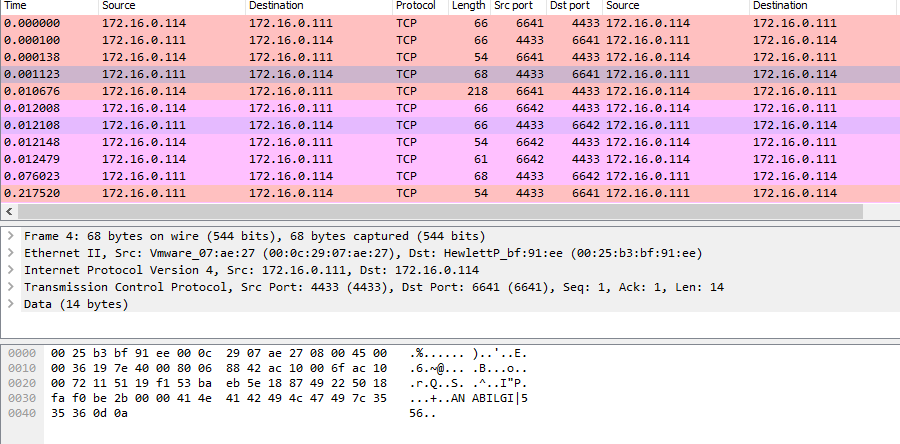
# Dataset:

For our project we used Remote Access Trojan (RAT) infected PCAP file:

[ratinfected.pcap](http://www.chrissanders.org/captures/ratinfected.pcap)

For Source location: [Click here](http://www.chrissanders.org/captures/)

This file has the presence of CYBEREXE RAT virus which can only be detected by the presence of the word ”ANABILGI” (Meaning: Basic information in Turkish language) in the 5-bytes information area of every packet. Below you can see the image of Wireshark where ANABILGI is present.



Since it is a PCAP file , we used Regular expression in Python to pull of the unstructured text into a structured information.