Steps

Corresponding ipynb:
https://github.com/shahifaqeer/oit-dns/blob/master/feature_extraction/pcap-feature-extractor.ipynb

- Load pcap and extract DNS and epoch time using tshark
    - 1 hour of data from midnight of 7th Feb 2016
- Load tshark csv and filter to A record only
- Group and key using srcip
- Extract the following features per source IP
    - number of destinations
    - number of DNS A queries
    - number of unique DNS A queries
    - time-diff for all domains: avg, 50%, min, max
    - most popular domain (1,2,3)
    - most popular domain (1,2,3) time-diff: avg
    - least popular domain
    - 1-Level domain and 2-Level domains
    - Frequency and Periodic features per domain (TODO)

Filters after feature extraction to list IoT devices

- A record only
- Filter out princeton.edu
    - Too many princeton.edu related domains in local data. This might be due to a specific capture point.
    - As we can't separate capture points it might be best to separate source IPs into dormnet and departments. We expect the capture point for dormnets to be outside campus, and thus should avoid repeated local queries
- Filter source IPs with number of unique 2LDs <= 10
- Filter source IPs with number of total domain queries >= 10
- The above filtering gives us 2223 devices of 9019 devices from 1 hour of data
- Save extracted feature table to processed/features_key_srcip.csv and combine with hostdb data using source IP as the key

TODO: Work on frequency/time period as features