ML Data Prep for User Action Classification

The ML will be trained using a single master CSV file, with all the extracted data from the training set of PCAP files. When calling the function that trains the ML, a percentage of the extracted data should retained to be used for testing.

Step 1: Refactor dataset files with type of action and sequence number

> Currently there are three folders of 50 files per action, with file names being the action type and a sequence 0-50. To merge each of the three folders for each action into one, a short python script *prepare_dataset.py* will need to be created, to save time manually modifying 500 files.

Step 2: Create a script ‘extract_features’ that iterates over a given PCAP file, extracts flows and features from each flow and returns a Pandas data frame of extracted data.

Data to be extracted:

1. Packet count
2. Average packet size
3. Standard deviation packet size
4. Average inter-arrival time (Time between packets arriving)

   The time between packets. Actions like "Search" often have a burst of small packets, while "Play" has sustained large downstream packets, therefore this metric can be used for classification.

5. Outbound ratio

   Ratio of incoming:outgoing data. Streaming a video will likely have a lot of incoming packets compared to posting a comment.

6. Target action
   a. What action was performed? One of: Like, Comment, Play, Search, Subscribe

Step 3: Create a method that iterates over each pcap file, passing the file to the extract_features function, and merging the result into one dataframe for export.

Output master.csv format example:

source_file	pkt_count	avg_pkt_size	std_pkt_size	avg_iat	outbound_ratio	target_action
like_01.pcap	450	1102.4	450.2	0.022	0.15	**Like**
like_02.pcap	425	1080.1	462.8	0.024	0.14	**Like**
search_01.pcap	890	640.5	310.1	0.011	0.35	**Search**
search_02.pcap	915	625.2	305.5	0.010	0.38	**Search**