TrackerIQ data science Home assignment

Intro

The TrackerIQ (TIQ) is a system to monitor application usage. The system is doing this by monitoring application logs. The system is extracting from the log events an action that the user is doing. The system is clustering user sessions and is able to detect anomalies by calculating the distance of the session from the cluster centroid. A session is considered as a histogram of all the Actions in the sessions. An action is constructed from key-value pairs from one log event.

Attached files.

Sessions_histogram.csv - CSV containing sessions and their histograms
Clusters_histogram.csv - CSV containing clusters and their centroid histogram
Sessions_distances.npy - The session distance from the cluster centroid matrix
Cluster_distances.npy - Cluster centroid distance matrix
data_sience.py - a Python example of loading the data and getting some stats from it

Challenge

What we ask for is a way to make sense of the sessions. The sessions you have are considered far from the centroids. We want a way to decide whether a session is “close” enough to a cluster that it can be compared with it. Or is it not related to any clusters and stands by itself? Can we find an underlying reason that will explain the distance? Like a specific action or key value?

Please provide

- Your conclusion in writing
- Try to describe the reasoning that led you to the conclusion
- Please attach any code you used.

Good luck!