DEVGRID

PYTHON SUL CONFERENCE SPRINT

This test exercises different aspects of your technical abilities when faced with requirements and expectations.

The result is a system capable of communicating, parsing and persisting events generated by an energy sensor. We expect the code to be concise, clear and organised. You are free to define services, interfaces, technologies, etc., that you feel are useful and/or necessary.

DEFINITION

You have installed an energy sensor in your house, and when you took a look at the events it was sending, this was the format you were faced with:

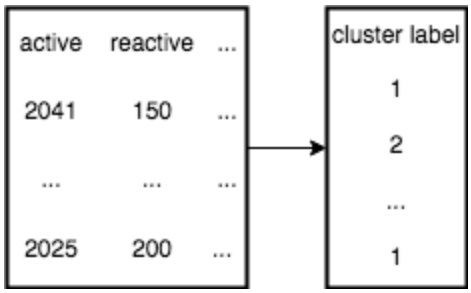
‘Device: ID=10; Fw=16071801; Evt=2; Alarms: CoilRevesed=OFF; Power: Active=289W; Reactive=279var; Appearent=403VA; Line: Current=1.75A; Voltage=230.08V; Phase=-43,841rad; Peaks: 1041.000; 1051.000; 1058.000; 1051.000; 1049.000; 1047.000; 1054.000; 1059.000; 1057.000; 1060.000; FFT Re: -257863.00; 102815.00; -64043.00; 48516.00; 59599.00; -4223.00; -43441.00; 23559.00; -24518.00; FFT Img: 481910.00; -14891.00; 69871.00; -7130.00; 43860.00; 34204.00; 55951.00; -6945.00; 26131.00; UTC Time: 2016-10-4 16:47:50; hz: 49.87; WiFi Strength: -62; Dummy: 20’

These events, although really dirty and unintuitive, provide useful electrical information that you want to use in a simple Machine Learning application that relies on what is called Clustering. Power includes the three different power values the sensor can measure, transients are also called peaks, and the harmonics, which are complex numbers, are calculated by fast Fourier transforms. Powers, transients, harmonics and all other measured values are vastly documented in the literature.

CLUSTERING

Clustering is the name given to a set of machine learning algorithms used to find unknown patterns in data. By applying a series of statistical techniques, these algorithms are able to identify similar data samples and label them accordingly. As the end result, a clustering

algorithm will return an array of labels, with the same size as the array of samples it was given, where each item will be the label to which the sample at the same index was assigned, as shown in the figure below.



YOUR TASK

You need to create web services that are capable of receiving the events one-by-one, parse them to a cleaner, standardised format, save them to a SQL database and, when the event count reaches 1000, the whole set of events should be sent in the format required by the `scikit-learn` library for its clustering algorithms, to feed the following code:

```
from sklearn.cluster import MeanShift, estimate_bandwidth
bandwidth = estimate_bandwidth(data, quantile=0.2, n_samples=200)
ms = MeanShift(bandwidth=bandwidth, cluster_all=False, bin_seeding=True)
labels = ms.fit_predict(data)
```

The variable `data` should contain only the active, reactive and apparent power values, the current and voltage, and the first three transients. Make sure that no other values are included.

You should get the set of labels returned by the `fit_predict()` method and save it in your SQL database, assigning the respective label to each event. This service should also be able to send the parsed, labeled data when requested.

One of the services you create should generate a report containing, at least:

- The number of events associated to each cluster
- The active power average for the events assigned to each cluster

Feel free to divide this task and create as many services as you think are appropriate.

SUGGESTIONS

- We like JSON
- Unit tests are useful to verify your implementation
- Use Python 3

DELIVERY

Your code must be hosted in a Git repository (GitHub, GitLab, Bitbucket, etc.). Include a succinct description on how to set up and execute your code.