Data Science Work Sample: Telematics

Root Insurance Co.

This task is intended to help us assess your capabilities as a data scientist. Your submission will be evaluated based on:

- Methodology
- Presentation style
- Programming ability (e.g. object-oriented design, variable naming, effective commenting, etc)
- Description of approach and implementation decisions
- Computational efficiency

Your work should be tracked locally on your computer with the version-control system `git`. Submit the tarball or zip file of the git repository via email when you are finished. Please do not share the data, instructions, or your implementation with any other party.

1 Task

The dataset consists of sensor data collected from a single user over many trips in their car. For each trip, data are obtained from two sources: the user's smartphone and an OBDII device. The task is to match trips from the smartphone to the corresponding trip obtained from the OBDII device. However, since the user may have taken smartphone trips in a different car or OBDII trips in their car without their smartphone, a one-to-one match is not expected to be possible.

2 Data

The prepared files include:

- `obd2_trips.json.gz` is a zipped json list containing vehicle speed data for each trip collected from the OBDII plug-in device.
- `mobile_trips.json.gz` is a zipped json list containing smartphone speed data for each trip collected from the phone.

The units for the datasets are as follows:

- accuracy: meters (lower value = more accurate)
- speed: unspecified
- timestamp: (epoch) seconds

3 Instructions

Implement a method (preferably in Python) that ingests both lists of trips and does the trip matching task described in Section (1). Some trips from each source may not have any match from the other source. Moreover, some trips from one source might match to multiple trips from the other source because trip start / trip end events differ by source.
Be sure to include instructions on how to run your code to both reproduce your results and produce matchings on a new set of data.

4 Output

An ideal submission will contain a comprehensive writeup of the methodology and findings. At the very least, we expect a list of matched tuples: $[(x_i, y_j), (x_k, y_m), ...]$ where x_i is the i-th smartphone trip and y_j is the j-th OBDII trip, each containing only the matched interval. The "matched interval" is the interval of time during which sensor data from both sources is present. Generally, this will be a proper subset of both trips.
A good way to present the output is to make a series of plots, each corresponding to a matched trip, with two lines plotted on it: the OBDII speed profile and the mobile speed profile. The time axes would be scaled to a common time so that the matched intervals (approximately) line up on the graph:

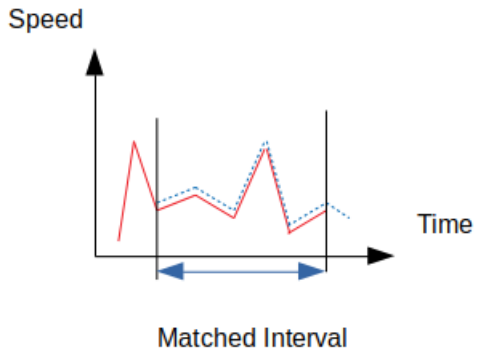

Figure 1: Matched Speed Profile

The code should be able to work on a dataset similar to this one, and should not be specialized to this particular set of trips.