1. Penjelasan dataset

Dalam tugas 2 ini, kami memilih dataset dengan judul “TAXI TRIP”. Dataset ini berisi data-data mengenai rekam jejak perjalanan dari taksi di suatu wilayah. Khususnya untuk data ini, diceritakan rekam jejak perjalanan beberapa taksi UBER pada suatu wilayah di Portugal.

Kumpulan data yang kami dapatkan mengandung atribut yang terdiri dari TRIP\_ID, CALL\_TYPE, ORIGIN\_CALL, ORIGIN\_STAND, TAXI\_ID, TIMESTAMP, DAY\_TYPE, MISSING\_DATA, dan POLYLINE. Setiap bagian menunjukkan informasinya masing-masing. Penjelasannya ialah sebagai berikut.

* TRIP\_ID menunjukkan ID dari setiap perjalanan
* ORIGIN\_CALL menunjukkan nomor telepon yang dihubungi dari permintaan customer
* ORIGIN\_STAND berisi sebuah ID yang dihubungkan dengan ID kota pada ‘metadata’ dan menunjukkan posisi awal taksi saat dihubungi *Customer*
* TAXI\_ID menunjukkan ID taxi yang telibat pada suatu perjalanan
* TIMESTAMP berisi informasi waktu dimana perjalanan dimulai
* DAYTYPE mengidentifikasikan keadaan hari pada perjalanan, misalnya pada saat liburan atau hari spesial lainnya (tipe B), sehari sebelum liburan (tipe C), atau hari-hari biasa dan weekend (tipe A)
* MISSING\_DATA menunjukkan apakah lokasi yang ditunjukkan oleh GPS hilang atau tidak. Jika tidak hilang, bernilai *False* dan sebaliknya.
* POLYLINE berisi koordinat-koordinat (dalam *latitude* dan *longitude*) yang menunjukkan *tracking* perjalanan taksi pada suatu trip. *Tracking* posisi taksi tersebut dilakukan setiap 15 detik sekali. Koordinat pertama menunjukkan posisi awal dari customer, koordinat akhir menunjukkan destinasi sebuah perjalanan.

1. Langkah Analisis

Pertama-tama, kita harus mengetahui informasi-informasi mana saja yang penting untuk dipakai dalam analisis data sesuai dengan kasus yang diberikan. Hal-hal yang diminta pada soal diantaranya.

1. Tempat yang paling sering dikunjungi pada trip yang ada
2. Lokasi akhir pada setiap perjalanan
3. Prediksi total waktu tempuh rata-rata pada setiap lintasan perjalanan

Maka, hal-hal tersebut membutuhkan informasi yang terdapat pada kolom ‘POLYLINE’. Kolom yang berisi informasi tentang *tracking* posisi taksi pada rentang waktu terssebut. Secara general, untuk mengakses *value* dari suatu atribut POLYLINE, kami menggunakan sebuah converter yang dimiliki *library* json, yaitu langsung mengubahnya menjadi *array of coordinates* sehingga lebih mudah untuk diakses.

* Kasus 1

Pada kasus ini, kami menambahkan suatu asumsi. Yakni, tempat yang dikunjungi pada setiap perjalanan ialah destinasi perjalanan yaitu koordinat terakhir pada POLYLINE. Karena jika dilihat, koordinat-koordinat lain yang terdapat pada sebuah data POLYLINE merupakan *tracking* taksi yang sedang bepergian sehingga bukan merupakan ‘tempat yang dikunjungi’.

Maka, tempat yang paling sering dikunjungi berarti titik atau koordinat akhir yang paling sering muncul dari seluruh data. Disini, kami menggunakan library collection yaitu Counter, yang dapat menentukan koordinat dengan kemunculan paling banyak pada keseluruhan data.

* Kasus 2

Untuk lokasi akhir, kita langsung saja ambil koordinat terakhir pada atribut POLYLINE. Untuk nama lokasinya, kita dapat mengakses file meta-data yang tersedia dan membandingkan koordinat *latitude* dan *longitude* nya. Untuk membandingkannya, kami menentukan sebuah titik apakah berada dalam rentang radiustertentu dari setiap nilai koordinat di meta-data.

* Kasus 3

Pada suatu perjalanan, ‘jarak’ antara dua koordinat nilainya 15 detik. Maka, kita cari untuk setiap data POLYLINE berapa panjang (len) array nya. Waktu tempuh total pada suatu perjalanan adalah (len-1)x15. Setelah semua waktu tempuh diakumulasikan dengan dijumlah, akhirnya dibagi jumlah data/ *tuple* yang ada. Hasilnya adalah rata-rata waktu tempuh. Namun masih ada kolom yang harus dipertimbangkan, yaitu pada MISSING\_DATA. Jika *value* nya TRUE, maka tidak akan sertakan dalam jumlah data (bagian pembagi). Karena jika *value* nya TRUE, maka lokasi tidak terdeteksi dan kami menganggap hal tersebut tidak termasuk dalam “perjalanan”.

1. Hasil Analisis