Roq.ad

# Opis projektu

## Cel projektu

Należy stworzyć algorytm, który będzie pozwalał na automatyczne znajdowanie grup identyfikatorów urządzeń, które należą do tego samego użytkownika.

## Dane

Dostępne są następujące trzy zbiory danych:

1. `learning-set` - zbiór uczący,
2. `verification-set` - zbiór weryfikujący,
3. `test-set` - zbiór testowy.

Zbiór uczący służy do ewentualnego uczenia tworzonego algorytmu, weryfikujący do weryfikowania jakości predykcji, natomiast zbior testowy służy do weryfikacji zaproponowanego rozwiązania przez Roq.ad.

Każdy z tych zbiorów danych zawiera trzy pliki typu CSV:

- `requests.csv` - poszczególne wiersze reprezentują jedno żądanie HTTP, które zostało wysłane z urzadzenia o danym identyfikatorze. Każdy wiersz zawiera dane o wysłanym żądaniu oraz informacje w jakich okolicznościach wystąpiło (np. timestamp, czyli czas wystapienia).
- `devices.csv` - każdy wiersz reprezentuje jedno urządzenie oraz składa się z atrybutów, które opisują to urządzenie, takie jak jego kategoria (np. czy jest to smartfon, czy może komputer desktopowy) lub nazwa używanego systemu operacyjnego.
- `labels.csv` - plik, który opisuje przynależność każdego urządzenia do użytkownika (każdy definiuje przynależność dla jednego urządzenia). Użytkownik może mieć jedno lub więcej urządzeń, natomiast urządzenie **może** należeć do kilku użytkowników.

Należy zauważyć, że plik `labels.csv` dla zbioru danych `test-set` jest pusty.

Opis atrybutów pliku `requests.csv`:

| Nazwa atrybutu | Opis |
|---|---|
| device_id | Zanonimizowany identyfikator urządzenia, które wygenerowało żądanie |
| ip | Zanonimizowane IP urządzenia, które wygenerowało żądanie |

| | |
|---|---|
| isp | Zanonimizowany identyfikator ISP urządzenia, które wygenerowało żądanie |
| country | Zanonimizowane państwo, z którego wysłano żądanie |
| region | Zanonimizowany region państwa, z którego wysłano żądanie |
| time_zone | Zanonimizowana strefa czasowa, z której wysłano żądanie |
| connection_type | Zanonimizowany typ połączenia, z którego wysłano żądanie |
| timestamp | Znacznik czasu (w milisekundach), w którym wysłano żądanie |
| url | Zanonimizowany znormalizowany url, z którego wysłano żądanie |

Opis atrybutów pliku `devices.csv`:

| Nazwa atrybutu | Opis |
|---|---|
| device_id | Zanonimizowany identyfikator urządzenia |
| os_name | Nazwa systemu operacyjnego urządzenia |
| os_version | Wersja systemu operacyjnego urządzenia |
| browser_name | Nazwa przeglądarki urządzenia |
| browser_version | Wersja przeglądarki urządzenia |
| device_name | Nazwa urządzenia (np. iPhone) |
| category | Kategoria urządzenia (np. DESKTOP, SMARTPHONE) |
| anonymous_1, anonymous_2, ... | Zanonimizowane cechy Roq.ad (wartości nominalne) |

Opis atrybutów pliku `labels.csv`:

| Nazwa atrybutu | Opis |
|---|---|
| device_id | Zanonimizowany identyfikator urządzenia |
| user_id | Zanonimizowany identyfikator użytkownika |

W plikach `requests.csv` oraz `devices.csv` mogą występować nieznane wartości atrybutów, które reprezentowane są przez wartość pustą.

## Ewaluacja wyników

Stworzony algorytm należy użyć do predykcji na danych ze zbioru testowego `test-set`, a następnie wygenerować plik wyjściowy CSV w którym wiersze reprezentują identyfikatory urządzeń, które należą do jednego użytkownika. W przypadku gdy urządzenie jest jedynym urzadzeniem użytkownika należy je również dodać do pliku.

Przykład:

```
dev_32,dev_4,dev_1
dev_5
dev_18,dev_101,dev_53
```