Location Request Analysis

A In Loco Media é a maior Ad Network baseada em geolocalização do mundo! Isso resulta em terabytes de dados de localização todo mês.

No link a seguir você encontrará uma amostra de dados real (embora simplificada em tamanho e atributos) dos nossos dados.

https://s3.amazonaws.com/ubee-public/data-samples/location_requests/north_america_sample.gz

Nesta amostra, no formato csv, encontram-se aproximadamente 20 milhões de respostas à requisições de localização. Cada linha consiste no seguinte *schema*:

> **"<mad_id>, <country>, <lat>, <lng>, <timestamp>, <source>"**

onde:

- **mad_id:** identificador único do usuário;
- **country:** sigla do país de origem da requisição. Pode ser "MX", "US" ou "CA";
- **lat:** latitude do ponto geográfico da requisição;
- **lng:** longitude do ponto geográfico da requisição;
- **timestamp:** timestamp do momento da requisição, formato: `"EEE MMM d HH:mm:ss zzz yyyy" (ex: "Thu Mar 24 14:08:08 BRT 2016")`
- **source:** tecnologia da rede pela qual localização foi identificada. Assume valores "gps" ou "wifi" nesta amostra.

Considerando os dados disponibilizados, escreva um Job por questão de forma que a saída do Job responda às perguntas.

Você pode utilizar tanto Hadoop quanto Spark (Java, Python ou Scala) para responder às seguintes questões:

1. Qual o número de usuários distintos por país? (DistinctUsersPerCountryJob)
2. Qual a hora do dia com mais requisições, em qualquer país? (PeakHourOfDayJob)
3. Qual o mad_id com mais requisições por país? (MostActiveUserPerCountryJob)
4. Quantas requisições são feitas em média por hora por país? (AverageHourlyRequestsCountPerCountryJob)

Para efeitos de organização, nomeie a classe principal de cada Job de acordo com o nome entre em parênteses ao final de cada questão.

Você deverá entregar um repositório no Github contendo as classes que descrevem a lógica de cada questão. Testes unitários não são necessários.

Indoor User Movement Prediction from RSS Data

Para esse problema, o seu trabalho será criar um classificador binário que irá responder à seguinte pergunta:

> **Dada uma série de passos por uma pessoa utilizando um dispositivo de *tracking* em seu corpo em um cômodo com dois ambientes diferentes, durante a movimentação houve uma mudança de ambiente?**

O dataset encontra-se disponível para download em: http://archive.ics.uci.edu/ml/datasets/Indoor+User+Movement+Prediction+from+RSS+data.

Leia cautelosamente o arquivo **dataset_description.txt.**

Caso alguma informação não esteja clara, não hesite em perguntar.

É permitido realizar o *split* de dados entre treinamento/teste ou treinamento/teste/validação da forma que você julgar melhor. Não é necessário implementar o classificador em si, sendo permitido o uso de bibliotecas *open source* para tal.

Você deverá entregar um repositório no Github contendo:

1. O código do projeto;
2. Instruções de como rodar o projeto no README;
3. Uma breve explanação concisa das suas principais decisões de projeto em um arquivo "REPORT.txt", tais como (mas não limitadas a):
    a. Qual algoritmo de classificação foi utilizado e o motivo;
    b. Qual a técnica de split utilizada e o motivo;
    c. Processo de *feature engineering*, caso haja;
    d. Descrição das métricas através das quais o modelo foi avaliado.