Trabalho Prático 1

2º semestre de 2019

I. Introdução

Nos últimos anos, a alta popularidade de plataformas de *microblogging* e redes sociais, como Twitter e Facebook, tem se tornado um meio-chave de comunicação tanto na *World Wide Web* (WWW). Por dia, usários de serviços como o Twitter, são capazes de gerar centenas de milhões de tweets. Aliado a isso, a era da Internet das Coisas (IoT), nos permite ter acesso a grande quantidade de dispositivos móveis capazes de produzir dados referenciados em localização e tempo.

A análise desses conteúdos é interessante em um grande número de aplicações, a soma de todos os textos de redes como o Twitter (tweets) pode, coletivamente, ser considerada uma fonte de informação sobre opiniões e sentimentos sobre produtos, política, sociedade e eventos. Como o Twitter é atualmente a plataforma de *microblogging* com o maior número de usuários ativos, diversos trabalhos vem sendo produzidos sobre a plataforma. Exemplos incluem previsão de receitas de bilheteria para filmes, flutuações dos mercados bolsistas, surtos de doenças como gripe e dengue, e mesmo eleições políticas.

Dados geo-referenciados têm especial importância pela oportunidade para vincular texto a locais e datas. A ideia de extrair padrões geo-temporais significativos a partir de textos de microblog relacionando a um contexto, permite a uma série de conhecimentos úteis para cientistas de dados de diversas áreas. Por exemplo, departamentos públicos de uma cidade, podem ter interesse em acompanhar um evento/show na cidade ou mapear surtos de alguma doença.

II. Objetivos

Neste trabalho prático, você está encarregado de analisar assuntos populares no Twitter. Você buscará analisar como eles se comportam espaço-temporalmente, por exemplo: Há determinados assuntos que ficaram restritos a uma determinada região (p.ex., delimitado por um raio)? Como é a popularidade desses assuntos pelo tempo? É possível encontrar algum evento que ocorreu na região? Use sua criatividade.

III. Base de dados

A base de dados consiste em extrações de tweets geo-referenciados do Twitter nos meses de março à junho de 2016, cada mês com aproximadamente 600 mil tweets e totalizando aproximadamente 2.4 milhões de tweets. O alvo foi a cidade de Curitiba e precisão da localização para esses tweets depende se usou-se GPS ou outra forma de localização. No pior caso, a precisão a ser considerada será de cidade.

O formato dos dados é o formato tradicional de um tweet[1], armazenado como um JSON por linha, contendo campos como: nome do usuário, id do usuário, data/hora de coleta do tweet, coordenadas (caso exista) de onde foi feito o tweet, o nome da cidade do tweet, o texto e as *hashtags* utilizadas no tweet. Devido a constante atualização da API do Twitter, pode haver

[1]Para mais informações: https://dev.twitter.com/overview/api/tweets

algumas diferenças entre os campos da listas na documentação e a base coletada. Por exemplo, na base coletada, a data/hora de envio de um determinado tweet podem ser verificados nos campos `created_at` e `timestamp_ms`, ambos representados em *Unix Time*. Em Python, o usuário pode converter-los em um formato convencional utilizando o comando `fromtimestamp` do modulo de `datetime`.

A base encontra-se disponível no cluster da disciplina em `hdfs:///datasets/geo_curitiba`, nessa pasta, todos os alunos tem acesso leitura. No entanto, cada aluno tem sua pasta individual em `hdfs://user/login`, onde poderá salvar seus resultados.

IV. Orientações

A identificação de tópicos no Twitter é área bastante estudada. Para essa disciplina, você poderá escolher qual estratégia utilizar poderá ser, por exemplo, a partir de um processamento de texto (no campo `text`) ou por *hashtags* (no campo `entities`).

Dependendo da precisão da localização, o geo-referenciamento de um tweet pode ser um ponto no formato (longitude, latitude) no campo `coordinates` ou um polígono, representando um perímetro, no campo `place`. Você é livre para decidir qual abordagem tomar, uma delas, seria converter o polígono pelo seu ponto central.

As coordenadas de latitude e longitude estão representadas em formato de graus decimais, popular em coordenadas de GPS. É interessante utilizar a formula de Haversine[2] para se calcular distâncias entre coordenadas, além do resultado ser em padrão de metros, leva-se em consideração a curvatura da Terra.

V. Documentação e Parâmetros de avaliação

Deverá ser escrito um relatório em que é explicado as análises, os resultados produzidos, como cada análise foi construída em Spark e quais suas premissas (caso exista). Não precisam colocar o código completo no relatório, apenas trechos para ajudar na sua ilustração. O aluno deverá enviar um único arquivo compactado contendo o(s) código(s)-fonte(s) e o relatório produzido. O aluno é livre para escolher qual linguagem utilizar (Scala, Python ou Java). Caso escolham Java, deverá ser enviado também as instruções para compilação.

Aproveite essa oportunidade para aprender a extrair informações úteis em grandes volumes de dados. Será avaliado a capacidade do aluno de interagir com o ambiente Spark bem como a qualidade de suas análises.

Obs.: Como o cluster é compartilhado para todos os alunos, cuidado para extrapolarem o armazenamento com múltiplos resultados intermediários. Para trabalhos como esse, as etapas de preprocessamento dos dados, ajudam na redução do tamanho final.

Boa sorte!

"The Answer to the Great Question... Of Life, the Universe and Everything... Is... Forty-two."
(The Hitchhiker's Guide to the Galaxy)

[2] https://bit.ly/2XWTUGz