Segundo exercício de programação

Covid-19

2º semestre de 2020

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, usuários de serviços como o Twitter são capazes de gerar centenas de milhões de *tweets*. A análise desses conteúdos é interessante para 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 vêm sendo produzidos sobre a plataforma. Mais recentemente, o Twitter vem sendo muito utilizado como ferramenta de divulgação e de coleta de informações sobre o Coronavirus (SARS-CoV-2), causador da pandemia global, e a doença que ele provoca (COVID-19). Nesse sentido, diversas iniciativas [1, 2, 3] estão buscando utilizar os *tweets* ao longo do mundo para tentar entender melhor alguns aspectos dessa pandemia.

II. Objetivos

Neste exercício de programação, você está encarregado de responder algumas perguntas sobre eventos relacionados ao Coronavirus durante a pandemia. Mais especificamente, você deverá responder:

1. Quais foram as 15 *hashtags* mais populares no mundo? E no Brasil?
2. Como foi a evolução de comentários sobre a pandemia ao longo das semanas nos Estados Unidos, Brasil e México (os três países mais afetados no continente americano[1] até dezembro/2019). Para responder essa pergunta, faça um único gráfico contabilizando a média móvel[2] do número de *tweets* ao longo dos dias, para cada país.
3. Houve usuários que se beneficiaram da pandemia para aumentar o número de seguidores? Mapeie os 100 usuários que mais cresceram em número de seguidores nesse intervalo da amostra. Quais desses usuários também estão na lista dos 100 usuário mais ativos?
4. Quais foram os 3 principais assuntos discutidos nos Estados Unidos? Extraia 5 palavras utilizadas em cada assunto (desconsiderando as *hashtags*).

[1] `https://news.google.com/covid19/map`

[2] A média móvel vem sendo um dos principais indicadores utilizados na pandemia, seu calculo é descrito em `https://g1.globo.com/bemestar/coronavirus/noticia/2020/07/27/entenda-como-e-calculada-a-media-movel-e-a-variacao-dos-casos-e-mortes-por-covid-19.ghtml`

III. Base de dados

A base de dados consiste em extrações de *tweets* públicos entre 29/03/2020 e 30/04/2020 a partir das principais *hashtags* relacionadas (`#coronavirus`, `#coronavirusoutbreak`, `#covid19`, `#covid_19`, `#coronavirusPandemic`, `#ihavecorona`, `#StayHomeStaySafe` e `#TestTraceIsolate`). Essa base é disponível na plataforma Kaggle[3] e consiste em aproximadamente 32 milhões de *tweets*, totalizando 4.9 GB.

A base disponibilizada, em arquivos *CSV* por dia, foi preprocessada e contém alguns campos do formato tradicional de um *tweet*[4], mais especificamente, os campos: *status_id*, *user_id*, *created_at*, *screen_name*, *text*, *source*, *reply_to_status_id*, *reply_to_user_id*, *reply_to_screen_name*, *is_quote*, *is_retweet*, *favourites_count*, *retweet_count*, *country_code*, *place_full_name*, *place_type*, *followers_count*, *friends_count*, *account_lang*, *account_created_at*, *verified*, e *lang*. O *dataset* já se encontra disponível no *cluster* da disciplina em `hdfs://compute1:9000/datasets/covid19/`, onde todos os alunos têm acesso de leitura. No entanto, cada aluno tem sua pasta individual em `hdfs://compute1:9000/user/<login>`, onde poderá salvar seus resultados, se necessário.

IV. Orientações

Um problema muito comum em cenários com grandes conjuntos de dados é a presença de elementos ausentes ou a má formatação da base. A base utilizada neste exercício também tem esses problemas. Por exemplo, por causa dos *tweets* poderem vir de diversas fontes como *Twitter Web App*, *Twitter for Android* ou até mesmo de serviços coletores, muitos campos podem ter sido processados com falhas ou informações vazias. Um exemplo disso é o campo *country_code*, em que por vezes pode conter o código `BR` ou `55` para representar o Brasil[5], outra vezes esse campo é nulo e a única informação relacionada é no campo *place_full_name*, que contém o nome completo da cidade, estado ou do país. Analise sua resposta para verificar suas premissas, tente tratar os erros durante a sua resposta para deixar sua solução mais fiel à realidade.

V. Documentação e Parâmetros de avaliação

Deverá ser escrito um relatório em que serão explicadas as análises, os resultados produzidos, como cada análise foi construída em Spark e quais suas premissas (caso existam). Não é necessário incluir 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, a entrega deve incluir também as instruções para compilação.

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

Obs. 1: Como o *cluster* é compartilhado para todos os alunos, cuidado para não extrapolar o armazenamento com múltiplos resultados intermediários. Para trabalhos como esse, as etapas de preprocessamento dos dados ajudam na redução do tamanho final.

[3] Disponível em `https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april` e `https://www.kaggle.com/smid80/coronavirus-covid19-tweets-late-april`

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

[5] Uma lista completa para outros países é disponível em `https://countrycode.org/`

Obs. 2: Lembrem-se de utilizar os parâmetros de execução de Spark para especificar os recursos de CPU/RAM desejados para executar suas aplicações em *cluster*.

Referências

[1] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273, May 2020.

[2] Christian E. Lopez, Malolan Vasu, and Caleb Gallemore. Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset. *arXiv preprint arXiv:2003.10359*, 2020.

[3] Sohaib R Rufai and Catey Bunce. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. *Journal of Public Health*, 42(3):510–516, 04 2020.

Boa sorte!

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