Bootcamp Analista de Dados

Desafio do módulo

Módulo 3	CDD – Coleta e Obtenção de Dados

Objetivos

- ✓ O objetivo principal deste trabalho é apresentar uma implementação em Python para a análise de sentimento e mineração de textos, envolvendo textos de tweets coletados via API e o pacote tweepy.

A coleta de tweets e a análise de polaridade (sentimento) utilizando o Python foi apresentada na aula 4.4.
Opcionalmente, os tweets podem ser coletados utilizando a linguagem R (aula 4.3) ou a plataforma Knime (aula 4.5), e depois submetidos à análise de polaridade via Python (pacote TextBlob).

Enunciado

O desafio consiste na uma implementação em Python para a análise de sentimento e mineração de textos, envolvendo textos de *tweets* coletados via API e o pacote *tweepy*.

Primeiramente, é necessário cadastrar uma conta no twitter e solicitar acesso de desenvolvedor. Depois, você deve criar sua aplicação no twitter e gerar as credenciais de acesso.

```
# Credenciais para utilização da API do Twitter

consumer_key = ""
consumer_secret = ""
access_token = ""
access_token_secret = ""
```

Os tweets coletados devem utilizar as seguintes palavras chave:

- ('home office' OR 'trabalho remoto' OR 'trabalho em casa' OR #homeoffice OR #trabalhoremoto OR #trabalhoemcasa) → Não é necessário informar a # na string.

Além disso, selecione os tipos de tweets “mixed” e inclua o valor total de tweets para o máximo possível. Para isso, defina o parâmetro count da função search do tweepy para 27000.

Exemplo de código em Python:

```
#Definir que palavra deseja pesquisar no Twitter

keyword = ('home office  OR  trabalho remoto  OR  trabalho em casa OR homeoffice OR  trabalhoremoto OR trabalhoemcasa')

# Fazer a busca por palavra chave
tweets = token.search(q=keyword,count=28000,result_type='mixed')
```

Atividades

Para isto, serão executadas as atividades:

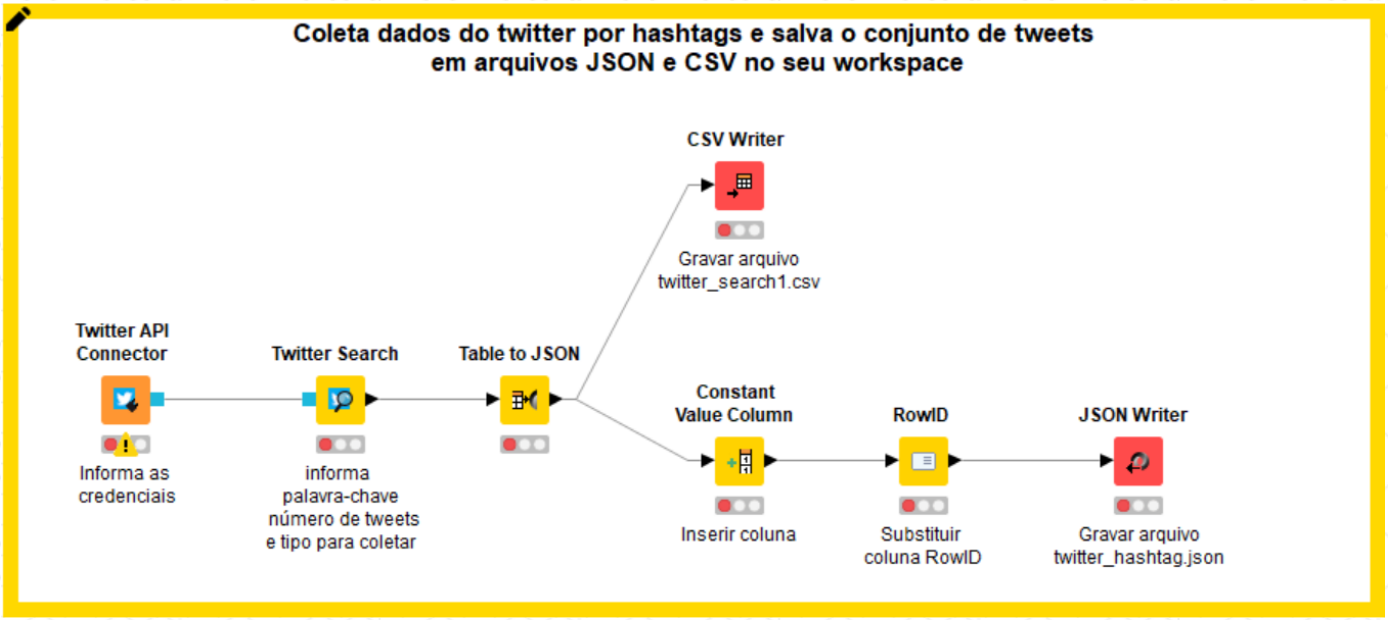
1º. Coleta de um conjunto de *tweets* através de API do Twitter utilizando o Python e seu pacote *tweepy*. Opcionalmente, os *tweets* podem ser coletados utilizando a linguagem R (aula 4.3) ou a plataforma Knime (4.5). Deve-se definir o parâmetro para quantidade de *tweets* coletados para 28000.

2º. Categorização dos *tweets* coletados, de forma que eles sejam identificados com sua respectiva polaridade, sendo uma *tweet* que represente sentimento positivo (polaridade > 0), negativo (polaridade < 0) ou neutro (polaridade = 0).

3º. Tokenização de palavras e definição da sua frequência conforme o sentimento que o *tweet* expressa, a partir dos termos coletados no texto dos *tweets* com sentimento positivo e negativo.

OBS: A mineração de texto será trabalhada na segunda aula interativa.

Para apoiar, segue um exemplo de fluxo no Knime:


Coleta dados do twitter por hashtags e salva o conjunto de tweets em arquivos JSON e CSV no seu workspace
CSV Writer
Gravar arquivo twitter_search1.csv
Twitter API Connector
Informa as credenciais
Twitter Search
informa palavra-chave número de tweets e tipo para coletar
Table to JSON
Constant Value Column
Inserir coluna
RowID
Substituir coluna RowID
JSON Writer
Gravar arquivo twitter_hashtag.json