

Análise de sentimentos do
Twitter
em temporeal com
Spark streaming e
NLTK



Análise de sentimentos do Twitter em tempo real com Spark Streaming e NLTK

- Ferramentas necessárias:

Python
(Anaconda)

NLTK
(Biblioteca de
Processamento de
Linguagem Natural)

Spark

API do Twitter
(conexão com Twiter)

Step1 – Simular Streaming de dados na porta 9898

Lembrar de abrir a porta com o comando [nc -lk 9898] no MobaXterm (no Windows)

Terminal: pyspark

1. Importar o StreamingContext
2. Criar um StreamingContext (sc) com intervalo batch de 1 segundo
3. Criar um Dstream(socketTextStream) que vai conectar na porta 9898 da sua máquina local (RECEIVER)
4. Criar um procedimento para ANÁLISE os dados em tempo real (AÇÃO E TRANSFORMAÇÃO)
5. Habilitar a leitura do Streaming de dados com o SparkStreaming

ssc.start() # Inicia a coleta e processamento do stream de dados

ssc.awaitTermination() # Aguarda a computação ser finalizada

Antes de dar o Stop (apertar no botão parar no JNB)

ssc.stop() # para o Streaming

RESUMO

1. Coleta
2. Aplicar as transformações no Dstreaming conforme nós definimos
3. Faz mapeamento e redução
4. Faz a contagem de palavras

Análise de sentimentos do Twitter em tempo real

- O Twitter é uma das redes sociais mais dinâmicas disponíveis atualmente. É possível coletar em tempo real informações preciosas e o sentimento das pessoas sobre os mais variados temas. São milhões de tweets por minuto em todo mundo e muita informação preciosa está escondida em cada tweet.
- O Streaming de dados gerado pelo Twitter pode alimentar aplicações analíticas, permitindo que empresas compreendam, em tempo real, o que clientes, parceiros e fornecedores estão pensando (e escrevendo) sobre você, sua marca, produto ou serviço
- Neste projeto, coletamos dados de Streaming do Twitter, aplicamos técnicas de análise em tempo real (à medida que os dados são gerados) e obtemos insights sobre um determinado assunto. O projeto pode ser facilmente reproduzido localmente em seu computador ou em um ambiente de cluster na nuvem. Tecnologias utilizadas:
 - **Python/NLTK** – construção do modelo de análise de sentimento
 - **Spark Streaming** – coleta do streaming de dados
 - **API do Twitter** – conexão a partir da nossa aplicação

Step2 – Análise de sentimentos do Twitter em tempo real

1. Preparar o SparkStreaming
2. Treinar o Classificador de Análise de Sentimentos (para identificar se o sentimento é positivo ou negativo) (usar conj de dados do Kaggle com 1.578.627 tweets)
 - 2.1 Realizar um processamento de linguagem natural (PNL) nos dados que serão usados para treinar o classificador.
 - 2.2 Coletar uma amostra de palavras que não são stopwords no dataset de treino (limpo)
 - 2.3 Treinar o algoritmo de ML (Classificador) para identificar a conotação de cada Tweet
3. Autenticação no Twitter
4. Configurar o Streaming de dados com o SparkStreaming

Processamento de Linguagem Natural nos dados que serão usados para treinar o classificador

1. Remover as stopwords
2. Converter as palavras em números
3. Treinar o classificador

Resultados

Zero: conotação negativa

Um: conotacao positiva

Horário	
	['10:41:24', [('0', 389), ('1', 112)]]
	['10:41:55', [('0', 371), ('1', 130)]]
	['10:42:26', [('1', 131), ('0', 370)]]
	['10:42:57', [('1', 138), ('0', 363)]]
	['10:43:29', [('1', 127), ('0', 374)]]
	['10:43:59', [('0', 375), ('1', 126)]]
	['10:44:30', [('1', 122), ('0', 379)]]
	['10:45:01', [('1', 131), ('0', 370)]]
	['10:45:33', [('0', 368), ('1', 133)]]

Os tweets foram descartados, apenas o resultado da classificação será salva (RDD ou .csv, por ex).

Assim, os tweets são coletados, processados, classificados e descartados.