Classificação de texto

com Hadoop + Spark Mlib

Erik Perillo, RA135582

Universidade Estadual de Campinas

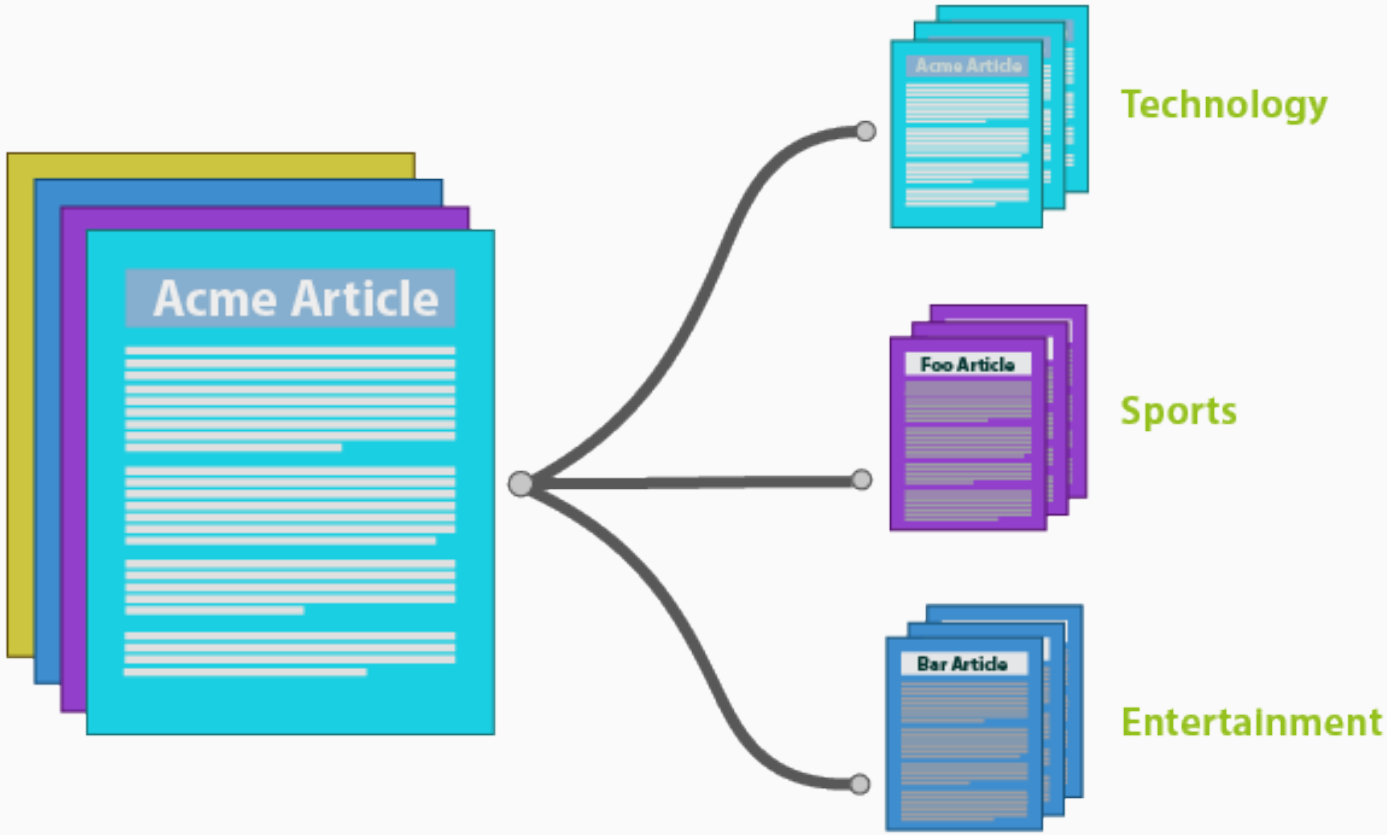
Objetivos

Objetivos

Um classificador de textos binário:

- Com uma das melhores técnicas atuais
- Com grande volume de dados

Classificação de Texto


Acme Article
Acme Article
Technology
Foo Article
Sports
Bar Article
Entertainment

Naive Bayes

$$p(C_k|x_1,\ldots,x_n) = \frac{1}{Z}p(C_k)\prod_{i=1}^{n}p(x_i|C_k)$$

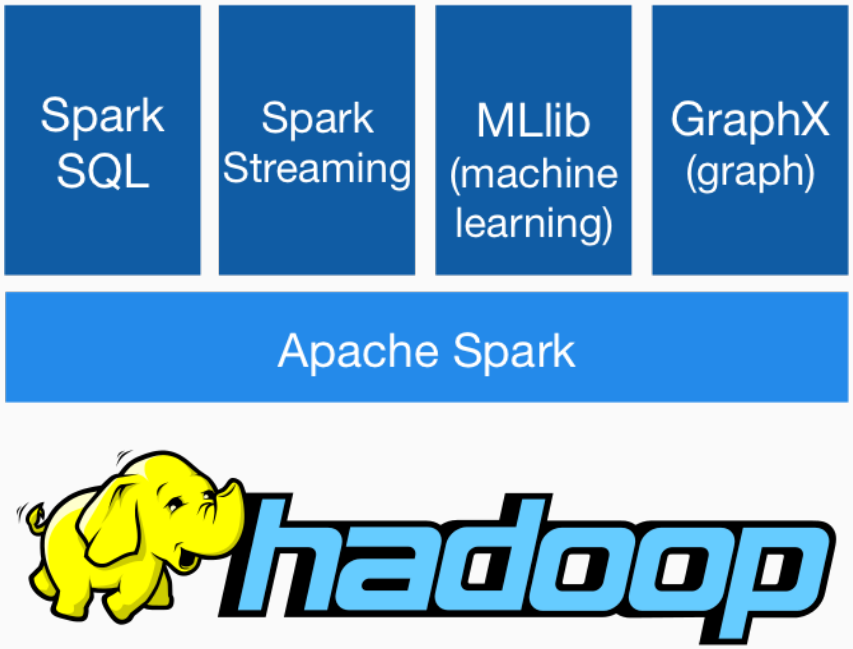
O projeto

Overview

Apache Spark

- Computação em *batches* de alto desempenho
- Análise de dados avançada
- Processamento de *stream* em tempo real
- Interfaces para Java, Scala, Python

Spark: componentes



Spark Mlib

- Diversas técnicas de *Machine Learning*
- Desenvolvimento rápido e escalável

Juntando tudo

Juntando tudo

- Tudo foi feito em Python :)
- Foi implementada visualização pela linha da comando

Flow dos dados

- Arquivos de uma database entram (`20_newsgroup`)
- Pré-processamento do texto é feito (*stemming*, *stopwords*...)
- Conversão de texto para vetor de *features* (`if-idf`)
- Textos positivos/negativos são convertidos para .csv
- Arquivo .csv é convertido em formato para *Spark*
- Treinamento é feito com *Naive Bayes*

Resultados