Spark MLlib

Machine Learning clusterizado

Esboço da Apresentação

- Um pouco sobre o Spark MLlib
- Fonte de dados
- Extração, transformação e seleção de características
- Classificação e Regressão
- Exemplo: Filtro de SPAM
- Exemplo: Analisador de sentimentos

Principais tópicos para discussão

Spark MLLIB

MLlib é uma biblioteca do Spark de aprendizado de máquina. Tem como objetivo ser escalável e fácil de usar. Fornece os recursos:

- Algorítmos de aprendizado de máquina: classificação, regressão, clusterização e filtragem colaborativa
- Pipelines: ferramentas para construção, avaliação e otimização
- Persistência: salvar e carregar algoritmos, modelos e pipelines
- Utilitários: algebra linear, estatisticas e etc

Fontes de Dados

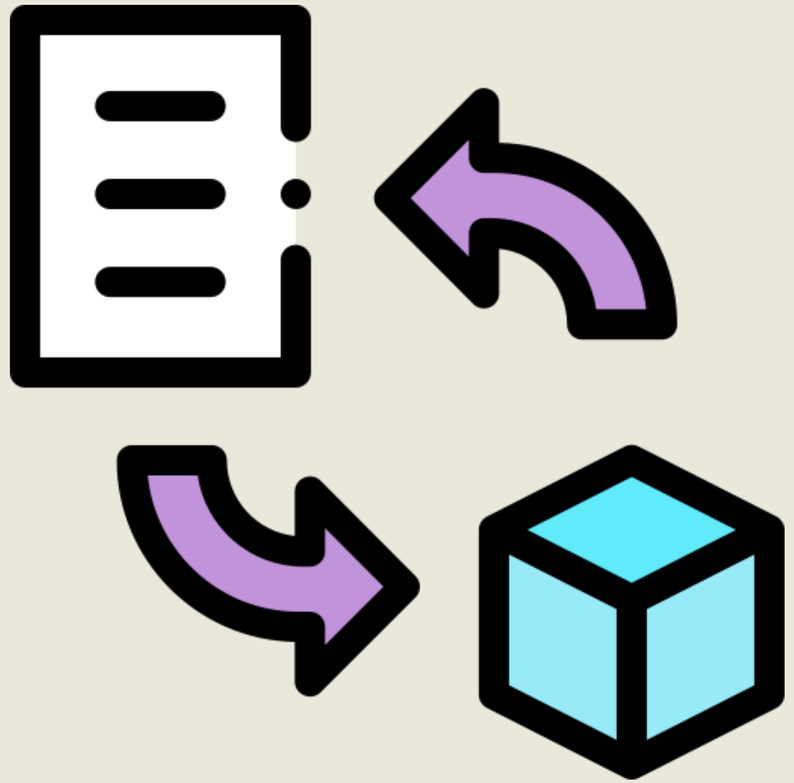
- Iniciando o processamento
- Entrada de dados padrão
  - *Parquet*
  - CSV
  - JSON
- Fontes Externas
  - Arquivos locais (file:///)
  - Hadoop (hdfs://)
  - Bancos de dados (PostgreSQL, Cassandra, entre outros)
- *Streaming* de Dados
  - Apache Kafka
  - Filas de mensageria

Extração de características

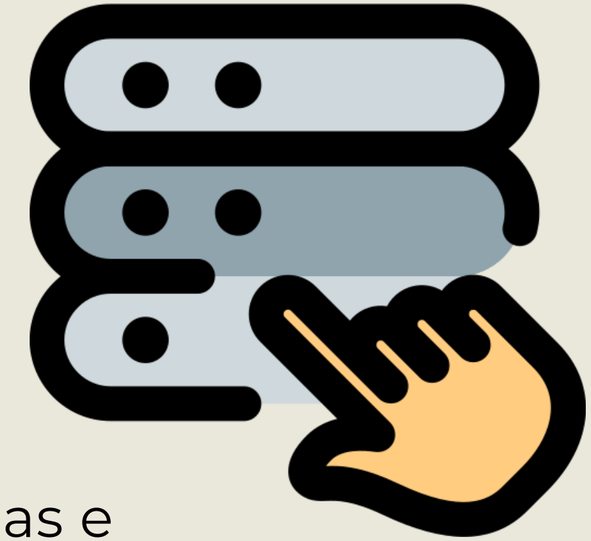
- TF-IDF
  - Term frequency–inverse document frequency (frequência do termo–inverso da frequência nos documentos)
  - Indica a importância de um termo baseado no número de ocorrência, mas equilibrando com o inverso da frequência
- Word2Vec
  - Mapeia as palavras em vetores de um único tamanho
  - Transforma o documento em vetor com a média de todas as palavras
  - Pode ser usado para predicação, similaridades de documento e etc
- CountVectorizer
  - Transforma o texto em um vetor com a contagem das palavras
- FeatureHasher
  - Transforma o texto em um vetor de hash

Transformação de características

- Tokenizer
- StopWordsRemover
- n-gram
- Binarizer
- PCA
- PolynomialExpansion
- Discrete Cosine Transform (DCT)
- StringIndexer
- IndexToString
- IndexToString
- OneHotEncoder
- VectorIndexer
- Interaction
- Normalizer
- StandardScaler
- RobustScaler
- MinMaxScaler
- MaxAbsScaler
- Bucketizer
- ElementwiseProduct
- SQLTransformer
- VectorAssembler
- VectorSizeHint
- QuantileDiscretizer
- Imputer



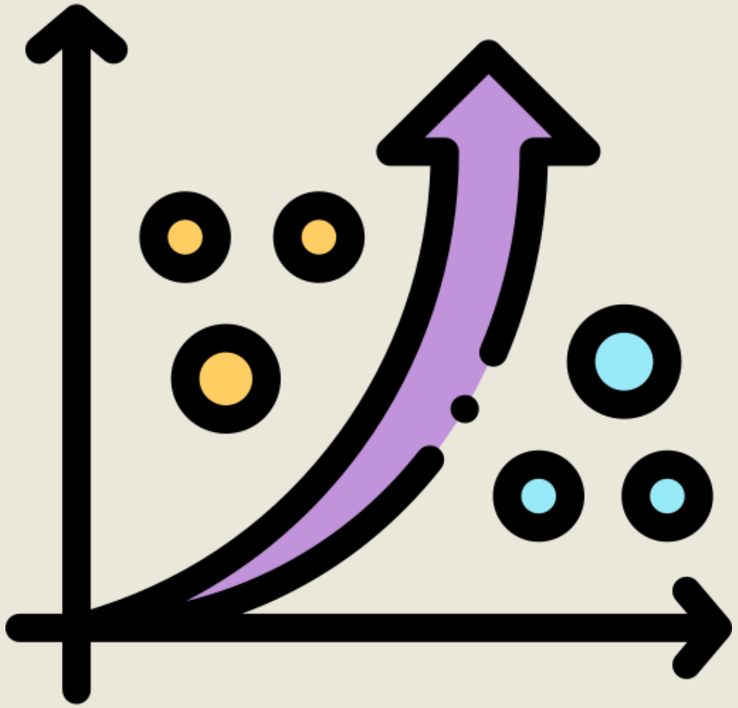
Seleção de características



- VectorSlicer
  - Transformador simples que recebe um vetor de colunas e retorna um outro vetor com as colunas selecionadas
  - Utiliza os índices ou nome das colunas para realizar a separação
- ChiSqSelector
  - Realiza a seleção pelo teste de independência Qui-quadrado ou X^2
  - Opera em dados rotulados com características selecionadas
- UnivariateFeatureSelector
  - Atua em dados categoricamente rotulados e características categoricamente selecionadas

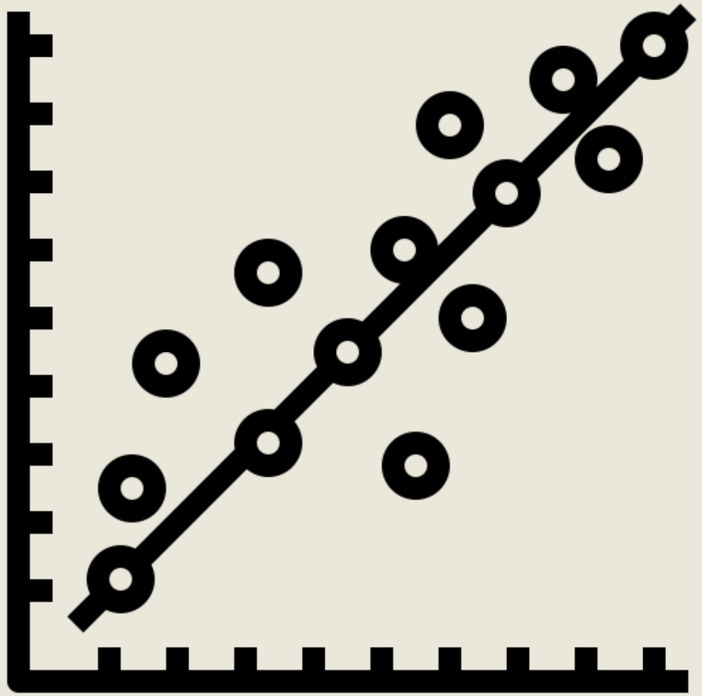
Classificação

- Logistic regression
  - Binomial logistic regression
  - Multinomial logistic regression
- Decision tree classifier
- Random forest classifier
- Gradient-boosted tree classifier
- Multilayer perceptron classifier
- Linear Support Vector Machine
- One-vs-Rest classifier (a.k.a. One-vs-All)
- Naive Bayes
- Factorization machines classifier



Regressão

- Regression
- Linear regression
- Generalized linear regression
- Available families
- Decision tree regression
- Random forest regression
- Gradient-boosted tree regression
- Survival regression
- Isotonic regression
- Factorization machines regressor

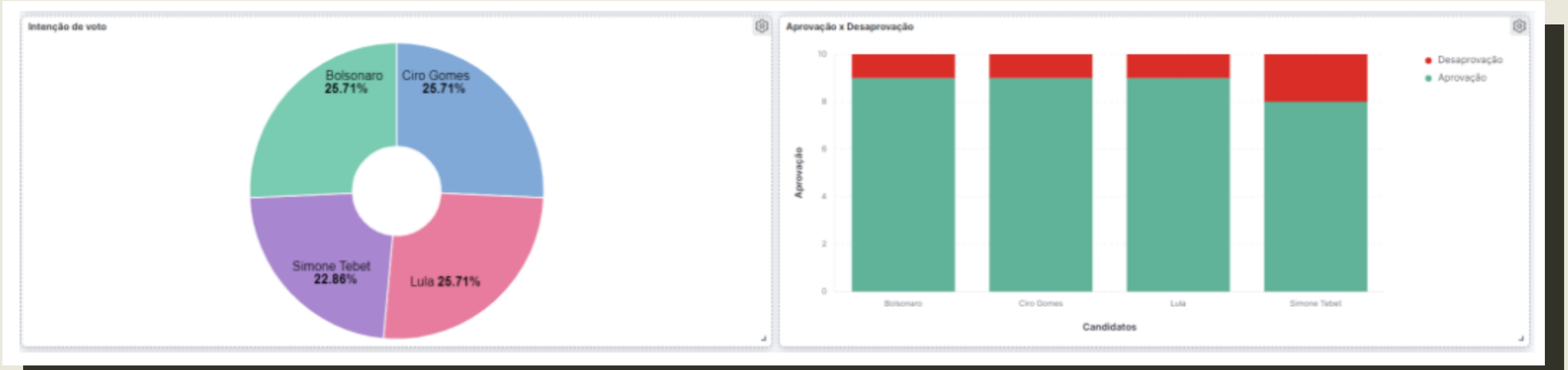


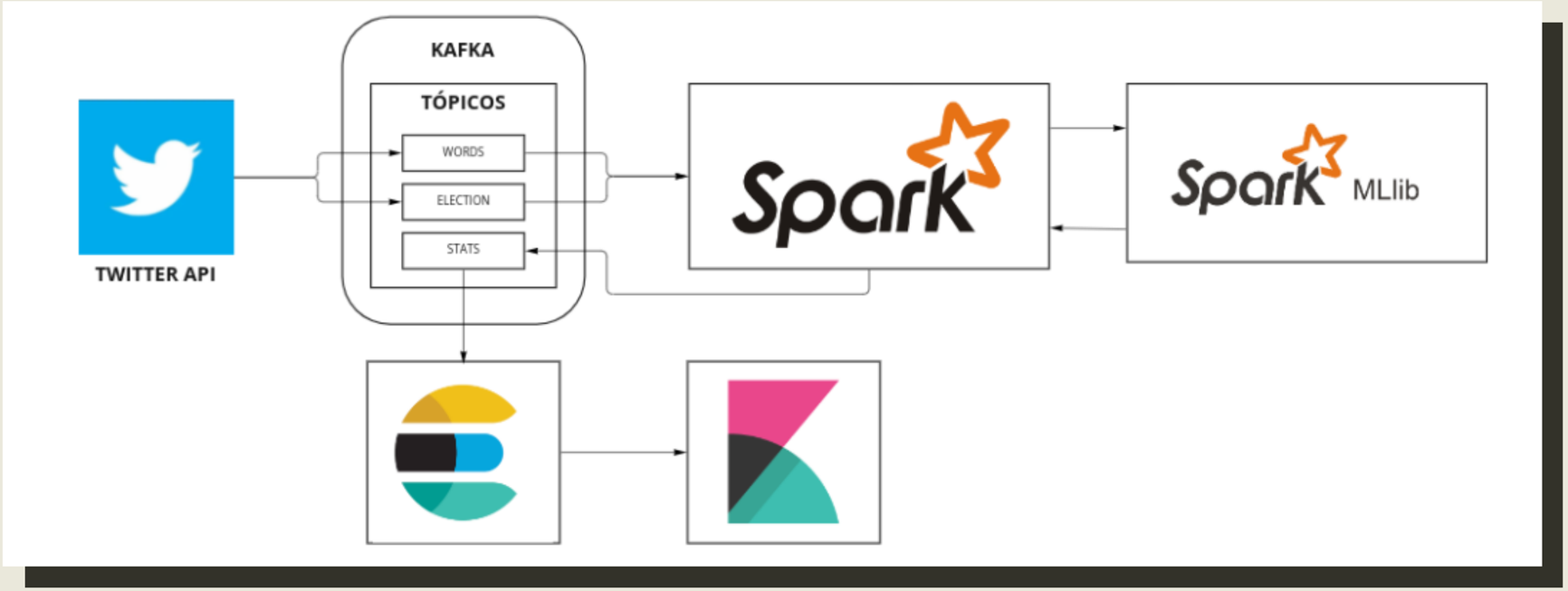
Exemplo 01

Filtro de SPAM

github.com/Joao-Moura/pspd_monitoria/tree/main/spark/mllib

Exemplo 02: Analisador de sentimentos na eleição de 2022





Repositório: github.com/thiagohdaqw/eleicoes