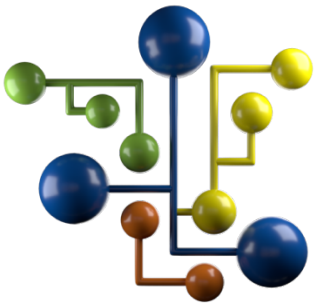

Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Principais Funcionalidades do Apache Spark MLLib

O MLLib possui também diversas funções para o trabalho de pré-processamento dos dados. Muitas dessas funções são encontradas no pacote mllib.feature.

Funcionalidade (Feature Extraction)	Funções (importadas a partir do pacote mllib.feature)
TF-IDF (Term Frequency – Inverse Document Frequency)	HashingTF() e IDF ()
Escala	StandardScaler()
Normalização	Normalizer()
Word2Vec	Word2Vec()
Estatística	colStats(), corr(), chiSqTest(), mean(), stdev(), sample()

Funções TF-IDF são usadas para tratamento de textos e geram vetores a partir de documentos de texto (como páginas web por exemplo). São computadas 2 estatísticas em cada termo em cada documento. O TF (Term Frequency) é o número de vezes que um termo ocorre no documento e o IDF (Inverse Document Frequency) é o número de vezes não frequentes (por isso o nome inverso) que um termo aparece em um documento, normalmente representado pelo Corpus (conjunto de dados de texto). Se multiplicarmos esses 2 índices (TF x IDF) teremos quão relevante é um termo em um documento específico. Normalmente processamos um texto com o pacote NLTK da linguagem Python antes de aplicarmos essas funções com MLLib. O NLTK é bem mais completo para atividades de mineração de texto, enquanto o MLLib é indicado para processamento paralelo e distribuído em cluster, por exemplo para tarefas de Processamento de Linguagem Natural ou Sistemas de Recomendação.

Muitos algoritmos de Machine Learning precisam que os dados estejam na mesma magnitude, ou seja, na mesma escala. Os algoritmos esperam por isso e vão funcionar corretamente se você, Cientista de Dados, entregar isso a eles. Um exemplo de escala é colocar os dados com a mesma média e desvio padrão igual a 1. Normalização dos dados é outra tarefa importante durante o pré-processamento e em MLLib, usamos a função Normalizer().

Word2Vec é um algoritmo baseado em redes neurais para textos, que pode ser usado com diversos outros algoritmos de análise. Similar ao TF/IDF.

Podemos também executar diversas operações e testes estatísticos com MLLib, com a vantagem de poder realizar o processamento em grandes conjuntos de dados.