Analytics e Inteligência Artificial

Aulas 1, 2 e 3
Databricks

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseadas em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

NOTA MÁXIMA ENADE 2019 MELHOR ESCOLA DO BRASIL
Único curso de graduação em administração a receber as notas máximas

QS WORLD UNIVERSITY RANKINGS
A primeira escola brasileira a ser finalista da maior competição de MBA do mundo

JOHN MOLSON SCHOOL OF BUSINESS
Única *Business School* brasileira a figurar no *ranking* LATAM

UN GLOBAL COMPACT
Signatária do Pacto Global da ONU

ANAMBA ASSOCIAÇÃO NACIONAL DE MBA
Membro fundador da ANAMBA - Associação Nacional MBAs

ASSOCIATION AMBA ACCREDITED
Credenciada pela AMBA - *Association of MBAs*

Executive MBA Council
Credenciada ao *Executive MBA Council*

AACSB Business Education Alliance Member
Filiada a AACSB - *Association to Advance Collegiate Schools of Business*

EFMD
Filiada a EFMD - *European Foundation for Management Development*

GUIA DO MBA ESTADÃO GUIA DO MBA SELECIONADO
Referência em cursos de MBA nas principais mídias de circulação

LABDATA FIA

NOSSOS DIFERENCIAIS | QUEM SOMOS

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.

Profª Drª Alessandra Martini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil

Os diretores foram professores de grandes especialistas do mercado

+10 anos de atuação

+1000 alunos formados

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

Corpo Diretivo

COORDENADORES DO LABDATA | ATUAÇÃO ACADÊMCIA E PROFISSIONAL

Profª Dra.
Alessandra Montini

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Têm muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em estatística aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Membro do Conselho Curador da FIA, Coordenadora de Grupos de Pesquisa no CNPQ, Parecerista da FAPESP e Colunista de grandes Portais de Tecnologia.

linkedin.com/in/alessandramontini/

Prof. Dr.
Adolpho Walter Canton

Diretor do LABDATA-FIA. Consultor em Projetos de *Analytics*, *Big Data* e Inteligência Artificial. Professor FEA – USP. PhD em Estatística Aplicada pela *University of North Carolina at Chapel Hill*, Estados Unidos.

Currículo – ProfJoão Nogueira

FORMAÇÃO ACADÊMICA | EXPERIÊNCIA PROFISSIONAL

- **(2019-Presente)** – Professor nos cursos de Extensão, Pós e MBA em Big Data e Data Mining na Fundação Instituto de Administração (FIA) – www.fia.com.br
- **(2021-Presente) –** Cientista de Dados na Cardif - https://bnpparibascardif.com.br
- **(2018-2020)** – Cientista de Dados na Via Varejo – https://viavarejo.com.br
- **(2016-Presente)** – Doutorando em Física Computacional e Estatística pelo Departamento de Física na Universidade Federal do Ceará – https://física.ufc.br
- **(2014-2016)** – Mestre em Física da Matéria Condensada pelo Departamento de Física na Universidade Federal do Ceará - https://física.ufc.br
- **(2012-2013)** – Estudante Intercambista na Universidade de Coimbra – Portugal – https://www.uc.pt
- **(2010-2014)** – Bacharel em Física pela Universidade Federal do Ceará – http://www.ufc.br
- **Contatos:**
  - E-mail: joaonogueira@fisica.ufc.br

Conteúdo da Aula

- 1. Introdução ao Databricks
- 2. Processamento de dados com PySpark no Databricks
- 3. Machine Learning com PySpark no Databricks

1. Introdução ao Databricks

1. Introdução ao Databricks

O Databricks é uma empresa fundada em 2013 pelos criadores do Spark. Construído sob uma moderna arquitetura na cloud, Databricks entrega uma combinação de data warehouses e data lakes para a criação de uma plataforma analítica unificada para processamento de dados e inteligência artificial.

1. Introdução ao Databricks

- A plataforma **Databricks** possibilita o processamento de dados utilizando o Spark, que é o principal framework open-source para processamento de grande volumes de dados.

- O Spark possui várias APIs, que são diferentes formas de interagir e executar comandos por meio de diferentes linguagens. Podemos executar comandos Spark dentro do Databricks com as seguintes linguagens de programação: **Python**, SQL, R, Java e Scala.

- A API do Spark que utiliza Python se chama **PySpark**. Podemos pensar no PySpark como se fosse um Pandas otimizado para lidar com qualquer volume de dados.

1. Introdução ao Databricks

Muitas companhias atualmente utilizam o Databricks, se beneficiando das suas múltiplas capacidades, como engenharia de dados em alta escala, ciência de dados colaborativa, desenvolvimento completo do ciclo analítico de machine learning e business analytics. Veja alguns exemplos.

O Databricks oferece uma conta gratuita, chamada **Databricks Community**, onde podemos aprender sobre a plataforma e utilizar muitos de seus recursos para projetos pessoais, como por exemplo desenvolver todo o ciclo de machine learning feito em aula para o problema de classificação.

1. Introdução ao Databricks

BENEFÍCIOS

Quais os benefícios do Databricks Community?

1. Gratuito;
2. Ambiente facilmente gerenciável para criação de clusters e aprendizagem de ferramentas de processamento de big data, como o Spark;
3. Suporte a diferentes APIs do Spark, como PySpark e Spark SQL;
4. Construído baseado no Jupyter.

1. Introdução ao Databricks

LIMITAÇÕES

Quais as limitações do Databricks Community?

1. Limite de inatividade: qualquer cluster criado será automaticamente desligado depois de um período de 2h de inatividade;
2. Limite de memória e processamento: é disponibilizado um cluster com 15GB de memória ram. Parece pouco, mas é suficiente para nossas atividades em classe.

1. Introdução ao Databricks

HANDS-ON

Cliquem no link abaixo para criarmos nossa conta e começar a brincadeira:

Tutorial – Utilização do Ambiente Databricks

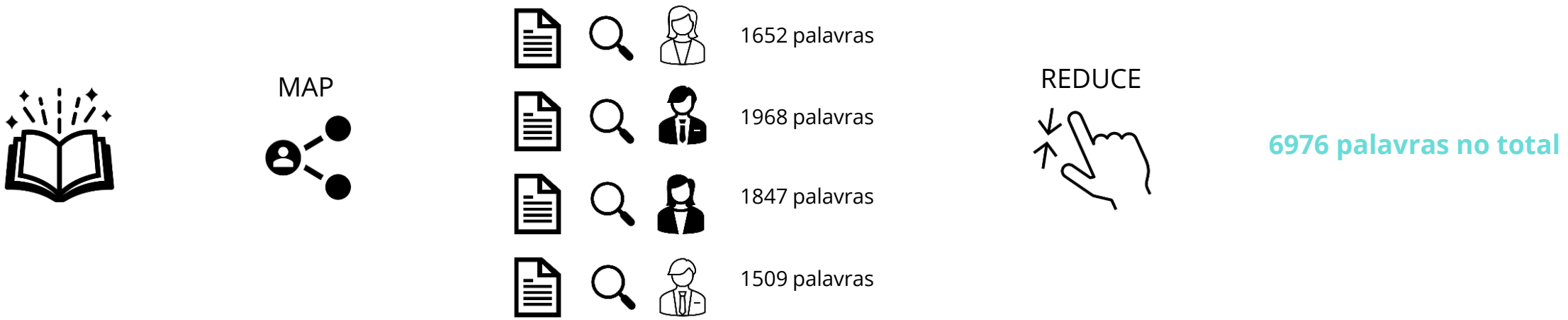
2. Processamento de Dados com PySpark no Databricks

1. Processamento de Dados com PySpark no Databricks

PROCESSAMENTO DE DADOS

PySpark é uma API para escrever código Spark utilizando a linguagem Python.

O Spark ficou bastante famoso por ser uma forma de processamento de Big Data superior ao **Hadoop Map Reduce**. Map Reduce é um modelo de framework para processar grandes volumes de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes. O diagrama abaixo ilustra de forma simples e intuitiva o processamento de big data com Map Reduce ao contar a quantidade de palavras em um livro:



1. Processamento de Dados com PySpark no Databricks

Apesar do Map Reduce ser bem intuitivo, ele é bastante lento pois exige bastante leitura e escrita dos dados em disco. Essa característica é onde o Spark é melhor, pois ele faz um intercâmbio de dados entre o disco e a memória ram, permitindo assim que as operações aconteçam de forma mais rápida pois boa parte do processamento acontece na memória ao invés de acontecer no disco, como é o caso para o Map Reduce. Além dessas características, Spark também possui um otimizador de queries, um DAG scheduler e uma physical execution engine. Não entraremos em detalhes a respeito dessas outras features aqui. Para nosso propósito, saber que o Spark realiza processamento na memória ram é o mais importante.

1. Processamento de Dados com PySpark no Databricks

Logo, podemos pensar no PySpark como uma versão poderosa do Pandas para lidar com grande volumes de dados. Apesar de serem projetos diferentes, cabe a analogia.

Em empresas que utilizam o Databricks, o PySpark é bastante utilizado para criação de ETLs e pipelines de dados. Dessa forma, podemos pensar no PySpark como a ferramenta para criar também as Analytical Base Tables, tabelas que são utilizadas para treinar modelos de machine learning.

Vamos por a mão na massa e recriar a ABT para o problema de classificação de propensão a não revenda para os sellers da Olist, como fizemos com Pandas nas aulas de aprendizagem supervisionada. Mas antes, vamos aprender alguns comandos básicos de manipulação de dados!

1. Processamento de Dados com PySpark no Databricks

Abram o seguinte notebook no Databricks:

1.0 Manipulação de Dados

1. Processamento de Dados com PySpark no Databricks

Agora que já aprendemos a manipular dados com PySpark, vamos recriar a nossa ABT de propensão a não revenda. O código se encontra no notebook:

2.0 Criando a ABT

3. Machine Learning com PySpark no Databricks

3. Machine Learning com PySpark no Databricks

Assim como utilizamos o PySpark para tratamento dos dados, podemos também utilizá-lo para treinarmos modelos de machine learning.

Benefícios

1. Código altamente escalável, sem importar o tamanho da ABT;
2. Manter a mesma linguagem tanto para criação da ABT quanto para treinamento do modelo;

Desvantagens

1. O módulo de machine learning do Spark (**mllib**) não possui tantas capacidades e flexibilidade quanto o scikit-learn;
2. Documentação ruim ou pelo menos não é tão boa quanto a do scikit-learn;
3. Ainda pouco utilizado pela comunidade, logo será difícil conseguir ajuda.

3. Machine Learning com PySpark no Databricks

HANDS-ON

Vamos treinar uma Random Forest para nos ajudar a prever a propensão de não revenda de um seller da Olist.

Abra o notebook no Databricks:

3.0 Treinando Modelo

Referências Bibliográficas

Referências Bibliográficas

LIVROS