Plano de Ensino

Inteligência na Web e Big Data

Fabricio Olivetti de França e Thiago Ferreira Covões
folivetti@ufabc.edu.br, thiago.covoes@ufabc.edu.br

Centro de Matemática, Computação e Cognição
Universidade Federal do ABC

UFABC

CMCC
Centro de Matemática, Computação e Cognição
Center for Mathematics, Computation and Cognition

Sobre os Docentes

prof. Fabricio Olivetti de França

Formado em Engenharia Elétrica com ênfase em Computação pela Universidade Católica de Santos.

Mestrado e Doutorado no mesmo curso na Universidade Estadual de Campinas.

Linhas de Pesquisa

- Co-agrupamento de Dados
- Regressão Simbólica
- Algoritmos Evolutivos
- Teoria das Categorias Aplicada
- Aprendizado de Máquina

prof. Thiago Ferreira Covões

Formado em Ciência da Computação pela Universidade Católica de Santos.

Mestrado e Doutorado no mesmo curso na Universidade de São Paulo - Campus São Carlos.

Linhas de Pesquisa

- Aprendizado de máquina
- Mineração de dados
- Algoritmos evolutivos

Sobre o Curso

Motivação

O avanço da tecnologia nos permitiu obter dados em massa de diversas fontes:

- Transações Bancárias
- Dados de medidas de sensores
- Experimentos genéticos
- Uso de Redes Sociais
- Construção de conteúdo colaborativo

Crescimento dos Dados

Grande quantidade de dados cria a necessidade de armazenamento.

- Maior HD tem cerca de 16TB por $\$\backslash 7.000,00$
- Dados na ordem de PetaBytes!!

Crescimento dos Dados

Grande quantidade de dados cria a necessidade de armazenamento.

- Maior HD tem cerca de 16TB por $\$\backslash 7.000,00$
- Dados na ordem de PetaBytes!!

Solução: Sistema de Arquivos Distribuídos

Crescimento dos Dados

Figura 1: Fonte: datacenterknowledge.com

Extração de Informação

Desses dados surge a necessidade de extrair informações úteis:

- Possíveis dívidas bancárias
- Genes relacionados a doenças
- Previsão de desastres naturais

Aprendizado de Máquina

- Muitos algoritmos consolidados
- Foco em bases de dados pequenas e bem estruturadas
- Bases de dados pequenas em apenas uma máquina!

Realidade

Temos:

- Bases de Dados grandes e distribuídas em diversas máquinas
- Dados brutos, não estruturados
- Sem a mínima ideia de como trabalhar com eles!

Objetivo do Curso

- Conhecimento básico sobre Ciência dos Dados
- Paradigma funcional / Pensamento Paralelo
- Enfoque em algoritmos escaláveis e distribuídos

Site da Disciplina

http://folivetti.github.io/teaching/2019-spring-teaching-2

Metodologia

- Aulas expositivas
- Atividades de Programação
- Projetos - Seminários

Cronograma

No site.

Material de apoio

Livros sobre Mineração de Dados:

Mining Massive Datasets

Introduction to Data Mining – P-N. Tan, M. Steinbach, V.Kumar – Addison Wesley 2005.

Data mining : pratical machine learning tools and techniques / 3. ed.

Material de apoio

Livros sobre Apache Spark:

https://www.gitbook.com/book/jaceklaskowski/mastering-apache-spark

ODERSKY, Martin et al. Programming in Scala. Mountain View, USA: Artima Press, 2008. xxxix, 736. ISBN 9780981531601.

Material de apoio

Livros sobre Python:

LUTZ, Mark; ASCHER, David; ALYH69. Aprendendo python. 2. ed. Porto Alegre, RS: Bookman, 2007. 566 p. ISBN 857780013x.

https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html

Avaliação

- 02 provas cada uma
- 01 projeto em grupo

Cada avaliação valerá 10 ptos e a média final será calculada via média harmônica.

Projeto

O projeto deve ser entregues via Github Classroom (link no site da disciplina)

Projeto

Cada aluno deve escolher um tema de projeto relacionado com a disciplina.

O projeto deve ser entregue em forma de relatório científico com até 3 páginas e códigos.

Apresentação de 10 a 15 minutos.

Projeto

Possíveis temas:

- Implementação distribuída de um algoritmo.
- Implementação de um algoritmo escalável.

Deve envolver uma quantidade suficiente de DADOS!

Conceito

O conceito final será calculado da seguinte forma:

```haskell
conceito :: Double -> Char
conceito nota
  | nota >= 9 = 'A'
  | nota >= 8 = 'B'
  | nota >= 6 = 'C'
  | nota >= 5 = 'D'
  | otherwise = 'F'
```