Escola Superior de Tecnologia de Tomar **Ano letivo:** 2022/2023

Mestrado em Engenharia Informática-Internet das Coisas
Mestrado, 2º Ciclo
Plano: PERA/2122/1500213 - 26/07/2022

Ficha da Unidade Curricular: Análise de Grande Volume de Dados
ECTS: 10; Horas - Totais: 260.0, Contacto e Tipologia, TP:30.0; PL:30.0; OT:30.0;
Ano | Semestre: 1 | S1
Tipo: Obrigatória; Interação: Presencial; Código: 390913
Área Científica: Ciências e Tecnologias da Programação

Docente Responsável
Ricardo Nuno Taborda Campos
Professor Adjunto

Docente(s)
Ricardo Nuno Taborda Campos
Professor Adjunto

Objetivos de Aprendizagem

1. Entender a importância do Python na Ciência de Dados
2. Ter conhecimento das questões éticas na coleta e no uso da informação
3. Dominar o processo de coleta de dados
4. Saber aplicar métodos de extração de informação
5. Estar familiarizado com os 5V's
6. Saber usar as principais frameworks

Objetivos de Aprendizagem (detalhado)

Esta UC tem por objetivo introduzir os alunos à extração e ao processamento de informação a partir de grandes volumes de dados. No final da UC o aluno deverá ter conhecimento das questões éticas associadas à coleta e ao uso de informação presente nos grandes conjuntos de dados; ser capaz de aplicar métodos de coleta e de extração de informação a partir da web com recurso à linguagem de programação Python; estar familiarizado com os 5V's dos grandes volumes de dados: volume, velocidade, variedade, veracidade e valor; saber usar e programar com recurso a frameworks de processamento de grandes volumes de dados a partir do

paradigma map-reduce.

Ao concluir esta UC o aluno deverá:
1. Entender a importância da linguagem de programação Python no contexto dos grandes volumes de dados
2. Ter conhecimento das questões éticas associadas à coleta e ao uso de informação presente nos grandes conjuntos de dados
3. Dominar o processo de coleta de dados a partir da web
4. Saber aplicar métodos de extração de informação
5. Estar familiarizado com os 5V's dos grandes volumes de dados: volume, velocidade, variedade, veracidade e valor.
6. Saber usar e programar as principais frameworks de armazenamento e processamento de grandes volumes de dados a partir do paradigma map-reduce.

Conteúdos Programáticos

1. Python no contexto dos Grandes Volumes de Dados
2. Ética e Privacidade dos Dados
3. Aquisição de Dados
4. Extração de Informação
5. Introdução ao Big Data (Modelo de Programação Map-Reduce)
6. Frameworks de Grandes Volumes de Dados

Conteúdos Programáticos (detalhado)

1. Python e o Big Data
1.1. Porquê usar Python?
1.2. História do Python
1.3. Características
1.4. Vantagens
1.5. Jupyter Anaconda

2. Ética e Privacidade dos Dados
2.1. Como é que podemos evitar o big data?
2.1. Questões éticas na coleta e uso de dados
2.2. Privacidade dos dados
2.3. Consequências do enviesamento (bias) do dados
2.4. Bolhas de Filtro (Filter Bubbles)
2.5. Fake News

3. Aquisição de Dados
3.1. Fontes e tipos de dados (estruturados; semi-estruturados; não estruturados)
3.2. Aquisição de dados a partir de Ficheiros (texto, imagens, pdfs, word, html, csv, json)
3.3. Aquisição de dados a partir da web (Bibliotecas Python; APIs; Web Scraping)

4. Extração de Informação
4.1. Definição

4.2. Arquitetura e pipeline de um sistema de extração de informação
4.3. Métodos de extração de informação

5. Introdução do Big Data (Modelo de Programação Map-Reduce)
5.1. O que é o big data?
5.2. Quem está a usar Big Data?
5.3. Origens da informação;
5.4. Razões para colecionar tantos dados;
5.5. Como é que o big data difere das tradicionais bases de dados?
5.6. 5 V?s do Big Data: volume, velocidade, variedade, veracidade e valor;
5.7. Diferentes Tipos de Processamento: Batch e Streaming

6. Frameworks de Armazenamento e Processamento de Grandes Volumes de Dados
6.1. Introdução ao Docker no contexto dos grandes volumes de dados
6.2. O paradigma Map-Reduce
6.3. Processamento Batch de grandes volumes de dados a partir de Hadoop e Dask
6.4. Processamento Streaming de grandes volumes de dados a partir de Spark

Metodologias de avaliação

Avaliação periódica: Projeto (60%) + Frequência (40%)

A entrega do projeto é obrigatória para a obtenção de aprovação na unidade curricular durante a avaliação periódica que pressupõe um mínimo de 70% de presenças. A entrega fora do prazo previsto implica a reprovação automática do aluno impossibilitando-o de se propor a exame. Os alunos ficam também automaticamente reprovados e excluídos de exame no caso de obterem nota inferior a 6 valores no projeto ou no caso de não atingirem um número mínimo de presenças.

Avaliação Final: Exame (100%)

Software utilizado em aula

Python: Anaconda e Jupyter Notebooks; PySpark

Estágio

Não aplicável

Bibliografia recomendada

- Santos, M. e Costa, C. (2019). *Big Data - Concepts, Warehousing, and Analytics* (pp. 1-312). FCA. Lisboa
- Foster, I. e Ghani, R. e Jarmin, R. e Kreuter, F. e Lane, J. (2017). *Big Data and Social Science. A Practical Guide for Methods and Tools* (pp. 1-349). Taylor & Francis. New York
- Erl, T. e Khattak, W. e Buhler, P. (2016). *Big Data Fundamentals: Concepts, Drivers & Techniques* . Prentice Hall. USA
- Sarkar, D. (2021). *Text Analytics with Python: A Practitioner's Guide to Natural Language*

Processing (pp. 1-661). Apress. USA

Coerência dos conteúdos programáticos com os objetivos

O programa cobre os diferentes objectivos e competências específicas que se pretendem proporcionar na unidade curricular, de acordo com a correspondência seguinte: Objectivos e competências/conteúdos programáticos
Objectivos 1: Conteúdos 1
Objectivos 2: Conteúdos 2
Objectivos 3: Conteúdos 3
Objectivos 4: Conteúdos 4
Objectivos 5: Conteúdos 5
Objectivos 6: Conteúdos 6

Metodologias de ensino

Exposição dos conteúdos programáticos com recurso ao método expositivo e demonstrativo. Análise e resolução de casos práticos através de notebooks. Os conhecimentos adquiridos serão avaliados através da realização e apresentação de projectos

Coerência das metodologias de ensino com os objetivos

Os objectivos de aprendizagem da unidade curricular são atingidos através do acompanhamento dos estudantes no
decurso da realização dos exercícios práticos e do projeto permitindo desta forma que os alunos solidifiquem as competências adquiridas.

Língua de ensino

Português

Pré-requisitos

Não aplicável

Programas Opcionais recomendados

Não aplicável

Observações

Objetivos de Desenvolvimento Sustentável:

4 - Garantir o acesso à educação inclusiva, de qualidade e equitativa, e promover oportunidades de aprendizagem ao longo da vida para todos;

Docente responsável

Ricardo Nuno Taborda Campos

Assinado de forma digital por Ricardo Nuno Taborda Campos

Homologado pelo C.T.C.
Acta n.º 35 Data 30/11/2022