

**Informática e Tecnologias Multimédia**

Licenciatura, 1º Ciclo

Plano: Despacho n.º 9184/2020 - 25/09/2020

**Ficha da Unidade Curricular: Técnicas Avançadas de Programação**

ECTS: 6; Horas - Totais: 162.0, Contacto e Tipologia, TP:28.0; PL:42.0; OT:5.0;

Ano | Semestre: 2 | S2

Tipo: Obrigatória; Interação: Presencial; Código: 814315

Área Científica: Informática

**Docente Responsável**

Ricardo Nuno Taborda Campos

Professor Adjunto

**Docente(s)**

Ricardo Nuno Taborda Campos

Professor Adjunto

**Objetivos de Aprendizagem**

O aluno deverá ser capaz de desenhar a arquitetura de um motor de busca (RI), explorar ferramentas de crawling e scraping, compreender as diferentes fases de processamento da linguagem natural e representação do texto, saber implementar um índice invertido, modelos de RI, e avaliação Cranfield

**Objetivos de Aprendizagem (detalhado)**

Esta unidade curricular tem por objectivo introduzir os alunos à recuperação de informação e motores de busca. Ao concluir esta unidade o estudante deverá:

- 1) ser capaz de desenhar a arquitetura de um motor de busca (RI);
- 2) ter conhecimento dos problemas relacionados com a ética e a privacidade dos dados;
- 3) saber explorar ferramentas de crawling e scraping;
- 4) compreender e implementar as diferentes fases de processamento da linguagem natural;
- 5) saber interpretar as características e estatísticas de um texto
- 6) saber representar um texto de acordo com as suas aracterísticas mais relevantes;

- 7) ser capaz de implementar um índice invertido;
- 8) ser capaz de implementar modelos de recuperação de informação;
- 9) compreender a metodologia de avaliação Cranfield

### **Conteúdos Programáticos**

- 1 - Introdução à Recuperação de Informação e Motores de Busca
- 2 - Ética e Privacidade dos dados na Recuperação de Informação
- 3 - Aquisição de Dados
- 4 - Pre-processamento de Texto
- 5 - Estatísticas do Texto
- 6 - Representação de Texto
- 7 - Índices Invertidos
- 8 - Modelos de Recuperação de Informação
- 9 - Avaliação de Sistema de Recuperação de Informação

### **Conteúdos Programáticos (detalhado)**

- 1 - Introdução à Recuperação de Informação e Motores de Busca
  - 1.1. Definição e objetivos
  - 1.2. Motores de busca
  - 1.3. Aplicações
  - 1.4. Dificuldades e desafios
  - 1.5. Arquitetura de um sistema de pesquisa de informação
- 2 - Ética e Privacidade dos dados na Recuperação de Informação
  - 2.1. Ética na recuperação de informação
  - 2.2. Privacidade dos dados
  - 2.3. Bias (enviesamento)
  - 2.4. Filter Bubbles
  - 2.5. Fake News
3. Aquisição de Dados
  - 3.1 Definição e objetivos
  - 3.2 APIs
  - 3.3 Web Scraping
  - 3.4 Web Crawling
  - 3.5 Web Dynamics
  - 3.6 Arquivos da Web
- 4 - Pre-processamento de Texto
  - 4.1 Parsing
  - 4.2 Divisão de frases
  - 4.3 Tokenização
  - 4.4 Stopwords
  - 4.5 Normalização
  - 4.6 Stemming

4.7 Reconhecimento de Entidades

4.8 Part of Speech

5 - Estatísticas do Texto

5.1 Lei de Zipf

5.2 Impacto na Recuperação de Informação

5.3 Co-ocorrência de palavras

6 - Representação do Texto

6.1 Características importantes do Texto

6.2 Vocabulário controlado vs Vocabulário livre

6.3 Bag of Words

6.4 Modelo de Espaço Vetorial

6.5 Importância dos termos

6.6 Matriz Documento - Termo

6.7 Matriz Termo - Documento

6.8 Introdução ao Topic Modeling e às Word Embeddings

7 - Índices Invertidos

7.1 Definição e objetivos

7.2 Estrutura de dados

7.3 Desafios

7.4 Big Data (Map Reduce)

7.5 Ferramentas de Indexação e de Pesquisa: Elasticsearch e PyTerrier

8 - Modelos de Recuperação de Informação

8.1 Algoritmo de RI

8.2 Modelo Booleano

8.3 Modelo de Espaço Vetorial

8.4 Modelo Probabilístico

8.5 Processamento de Queries

9 - Avaliação de Sistema de Recuperação de Informação

9.1 Noção de Relevância

9.1 Métodos de Avaliação

9.2 Coleções de Teste

9.3 Métricas de Avaliação

9.4 Testes de Significância

### **Metodologias de avaliação**

Avaliação por frequência: [Projeto I (20%) + Projeto II (10%) + Projeto III (10%)] + Projeto Final (30%) + Frequência (30%) [prova com consulta]. Os alunos deverão ter, em cada um dos elementos de avaliação, uma nota mínima de 6 valores. A classificação final da UC resulta da média ponderada das classificações obtidas nas componentes de avaliação definidas. O aluno obtém aprovação à UC, estando dispensado de Exame, de acordo com o disposto nos Pontos 11 e 12, do Artigo 11º, do regulamento Académico do IPT.

Avaliação por exame: Projeto (50%) + Exame (50%) [prova com consulta parcial dos conteúdos]. Os alunos deverão ter, em cada um dos elementos de avaliação, uma nota mínima de 6 valores. A classificação final da UC resulta da média ponderada das classificações obtidas nas componentes de avaliação definidas. O aluno obtém aprovação à UC, estando dispensado de Exame, de acordo com o disposto nos Pontos 11 e 12, do Artigo 11º, do regulamento Académico do IPT.

Requisitos de admissibilidade à frequência e ao exame:

- Mínimo de 70% de assiduidade às aulas (exceto trabalhadores estudantes);
- As presenças em aula não são classificadas com nota nem contam para avaliação, constituem, no entanto, condição necessária para aprovação à UC por frequência e exame. O incumprimento deste item impede o aluno de se submeter à frequência e ao exame.

### **Software utilizado em aula**

Python - Anaconda

Jupyter Notebooks

PyCharm Community

Moodle: plataforma de eLearning do IPT, Centro de eLearning

### **Estágio**

Não Aplicável

### **Bibliografia recomendada**

- Baeza-Yates, R. e Ribeiro-Neto, B. (2010). *Modern Information Retrieval* (pp. 1-944). 1, Addison Wesley Longman Publishing Co. Inc.. USA
- Gomes, D. e Demidova, E. e Winters, J. e Risse, T. (2021). *The Past Web. Exploring Web Archives* (pp. 1-297). Springer. Lisboa
- Croft, B. e Metzler, D. e Strohman, T. (0). *Search Engines: Information Retrieval in Practice* Acedido em 12 de fevereiro de 2019 em <http://ciir.cs.umass.edu/irbook/>

### **Coerência dos conteúdos programáticos com os objetivos**

Os conteúdos programáticos estão em coerência com os objetivos da unidade curricular, atendendo a que:

- O ponto 1 dos conteúdos programáticos pretende concretizar o ponto 1 dos objetivos
- O ponto 2 dos conteúdos programáticos pretende concretizar o ponto 2 dos objetivos
- O ponto 3 dos conteúdos programáticos pretende concretizar o ponto 3 dos objetivos
- O ponto 4 dos conteúdos programáticos pretende concretizar o ponto 4 dos objetivos
- O ponto 5 dos conteúdos programáticos pretende concretizar o ponto 5 dos objetivos
- O ponto 6 dos conteúdos programáticos pretende concretizar o ponto 6 dos objetivos
- O ponto 7 dos conteúdos programáticos pretende concretizar o ponto 7 dos objetivos
- O ponto 8 dos conteúdos programáticos pretende concretizar o ponto 8 dos objetivos
- O ponto 9 dos conteúdos programáticos pretende concretizar o ponto 9 dos objetivos

## **Metodologias de ensino**

Aulas teórico-práticas expositivas onde se descrevem os conceitos fundamentais. Aulas práticas-laboratoriais de resolução de casos práticos e aplicação dos conceitos a cenários de utilização real.

## **Coerência das metodologias de ensino com os objetivos**

Os objetivos de aprendizagem do curso são atingidos através da realização de um conjunto de exercícios práticos permitindo desta forma que os alunos solidifiquem as competências adquiridas. Considera-se ainda importante a orientação tutorial, onde o docente procura esclarecer dúvidas e apontar soluções para o sucesso do processo de aprendizagem da UC.

## **Língua de ensino**

Português

## **Pré-requisitos**

Preferencialmente o aluno deverá ter tido aproveitamento à UC de Linguagens de Programação

## **Programas Opcionais recomendados**

Não aplicável.

## **Observações**

Os conteúdos da UC serão trabalhados tendo em vista o cumprimento dos objetivos de desenvolvimento sustentável (ODS)

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  
Dados: 2022.02.25  
13:05:08 Z

---