



# Processamento e Recuperação de Informação

## Apresentação da Disciplina

Departamento de Engenharia Informática  
Instituto Superior Técnico

1º Semestre  
2018/2019



# Apresentação (1)

Processamento  
e Recuperação  
de Informação

## Professores



Pável Calado (responsável)



Bruno Martins



João Monteiro



Danielle Vieira



# Apresentação (1)

Processamento  
e Recuperação  
de Informação

## Professores



Pável Calado (responsável)



Bruno Martins



João Monteiro



Danielle Vieira



# Apresentação (2)

Processamento  
e Recuperação  
de Informação

## Tema da Disciplina

Busca, extração e análise de informação expressa textualmente, e.g. existente na World Wide Web.

## Aulas

- Teóricas: Conceitos Fundamentais + Teoria + Exemplos
- Laboratório: Problemas Práticos + Exercícios + Apoio ao Projeto



# O que vão aprender...

- Projetar soluções modernas para o processamento, gestão e interrogação de **grandes volumes de informação não estruturada**;
- **Classificar e agrupar automaticamente** conjuntos de recursos (e.g., grandes conjuntos de documentos de texto) através de características descritivas;
- Conceber sistemas para a **recuperação e filtragem da informação** relevante existente em grandes coleções, com base em termos chave, com base em exemplos, ou com base em perfis dos utilizadores;
- Conceber sistemas para a **extração de informação** a partir de documentos textuais ou da Web;
- **Avaliar comparativamente diferentes sistemas** para a extração, filtragem e recuperação de informação relevante.



# Material de Apoio

Processamento  
e Recuperação  
de Informação

## Bibliografia Principal

Ricardo Baeza-Yates and Berthier Ribeiro-Neto, **Modern Information Retrieval**, 2ª ed. (2011)

<http://www.mir2ed.org>

Bing Liu, **Web Data Mining**, 2ª ed. (2011)

<http://www.cs.uic.edu/~liub/WebMiningBook.html>

## Bibliografia Secundária

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, **Introduction to Information Retrieval** (2008)

<http://nlp.stanford.edu/IR-book/>

Anand Rajaraman, Jure Leskovec and Jeffrey D. Ullman, **Mining of Massive Datasets** (2013)

<http://infolab.stanford.edu/~ullman/mmds.html>

Ian H. Witten, Alistair Moffat, Timothy C. Bell, **Managing Gigabytes: Compressing and Indexing Documents and Images**, 2ª ed. (2000)

<http://people.eng.unimelb.edu.au/ammoffat/mg/>

Outras referências serão disponibilizadas ao longo do semestre.



# Avaliação

Processamento  
e Recuperação  
de Informação

- Exame (individual) = 70%; nota mínima = 9.5
  - Exame com consulta, mas limitada a **uma folha A4 manuscrita**.
- Projeto (em grupo) = 30%; nota mínima = 9.5
  - Entrega de relatório + apresentação final do projeto
  - Grupos de **3 alunos**.



# Trabalhadores-Estudantes e Época Especial

Processamento  
e Recuperação  
de Informação

## Avaliação para Trabalhadores-Estudantes

- Mesmo método de avaliação;
- Alternativamente, alunos podem optar por apenas fazer o exame.

Quem fizer projeto em grupo será avaliado como aluno regular

## Avaliação em Época Especial

- Avaliação com base num exame.





# Datas para Avaliação

Processamento  
e Recuperação  
de Informação

- Projeto: 07/12/2018
  - Apresentações na semana de 10/12
- Exame 1 : 8/01/2019 - 11h30
- Exame 2 : 5/02/2019 - 8h00



# Programa

Processamento  
e Recuperação  
de Informação

- 1 Introdução à extração e recuperação de informação
- 2 Modelos clássicos de recuperação de informação
- 3 Classificação e agrupamento de documentos
- 4 Informação não estruturada e extração de informação textual
- 5 Avaliação em recuperação e extração de informação
- 6 Análise de hiperligações e recuperação de informação na Web
- 7 Aprendizagem automática para ordenação de documentos
- 8 Extração de documentos da Web
- 9 Pesquisa por similaridade em dados multi-dimensionais
- 10 Implementação de sistemas de recuperação de informação
- 11 Aplicações



## Linguagem de programação: **Python**

### Recomendações:

- Comecem a praticar **hoje!**
- Formem os grupos **o mais depressa possível**
- Usem os vossos portáteis nas aulas de lab., se possível



# Python

Processamento  
e Recuperação  
de Informação

## Para começar:

Python Programming Language <http://www.python.org/>

The Python Tutorial <http://docs.python.org/tutorial/>

The Python Standard Library <http://docs.python.org/library/>

Python Tutorial @ w3schools <https://www.w3schools.com/python/>

## Outras ferramentas úteis:

Natural Language Toolkit <http://nltk.org/>

Whoosh <http://pypi.python.org/pypi/Whoosh/>

Beautiful Soup <http://www.crummy.com/software/BeautifulSoup/>

feedparser <https://github.com/kurtmckee/feedparser>

NumPy <http://www.numpy.org/>

scikit-learn <http://scikit-learn.org/>

... e outras a ser apresentadas ao longo das aulas



Mais questões?