
Ano Letivo 2020-21

Unidade Curricular PROCESSAMENTO DE GRANDES DADOS

Cursos ENGENHARIA INFORMÁTICA (2.º ciclo) (*)

(*) Curso onde a unidade curricular é opcional

Unidade Orgânica Faculdade de Ciências e Tecnologia

Código da Unidade Curricular 14741055

Área Científica CIÊNCIAS DE COMPUTAÇÃO

Sigla

Línguas de Aprendizagem Português (PT) ou Inglês (ENG)

Modalidade de ensino On-line

Docente Responsável António Eduardo de Barros Ruano

DOCENTE	TIPO DE AULA	TURMAS	TOTAL HORAS DE CONTACTO (*)
António Eduardo de Barros Ruano	PL; T	T1; PL1	28T; 28PL

* Para turmas lecionadas conjuntamente, apenas é contabilizada a carga horária de uma delas.

ANO	PERÍODO DE FUNCIONAMENTO*	HORAS DE CONTACTO	HORAS TOTAIS DE TRABALHO	ECTS
1º	S2	28T; 28PL	156	6

* A-Anual;S-Semestral;Q-Quadrimestral;T-Trimestral

Precedências

Sem precedências

Conhecimentos Prévios recomendados

Redes Neurais e Aprendizagem Profunda

Objetivos de aprendizagem (conhecimentos, aptidões e competências)

O aluno, no final do curso, deve ser capaz de:

- Processar diferentes tipos de dados (grande dimensão, grafos, infinitos, rotulados)
- Usar diferentes modelos de computação (MapReduce, fluxos e algoritmos online)
- Resolver problemas reais (sistemas de recomendação, análise de cesta de compras, deteção de spam, deteção de documentos duplicados)

Conteúdos programáticos

1. Introdução - o que é o processamento de grandes dados?
2. Programação Hadoop / MapReduce
3. Encontrando Itens Similares
4. Fluxos de Dados de Mineração
5. Análise de Links
6. Conjuntos de itens frequentes
7. Agrupamentos
8. Publicidade na Web
9. Sistemas de Recomendação
10. Mineração de gráficos de redes sociais
11. Redução de Dimensionalidade
12. Aprendizagem Máquina em Grande Escala

Metodologias de ensino (avaliação incluída)

As aulas serão distribuídas por componentes Teórica (T) e Prática Laboratorial (PL). Nas aulas T será exposta a teoria, e apresentados exemplos de aplicação. Nas aulas PL os estudantes implementarão programas em Hadoop, R ou Python, utilizando os módulos disponíveis em cada uma destas linguagens/ambientes.

Nesta disciplina haverá um mini-projeto em grupo, e um trabalho individual, ambos com um peso de 50% para a nota final. Se a nota final for igual ou superior a 10, e ambas as componentes forem classificadas com uma nota superior a 8, o aluno passa.

Bibliografia principal

Apontamentos on-line

Marz, N. and J. Warren, Big Data: Principles and best practices of scalable realtime data systems. 2015: Manning Publications.

Hwang, K. and M. Chen, Big-Data Analytics for Cloud, IoT and Cognitive Computing. 2017.

Ratner, B., Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. 3 ed. 2017: CRC Press.

Lin, J., & Dyer, C. (2010). Data-Intensive Text Processing with MapReduce: Morgan and Claypool Publishers.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of Massive Datasets: Cambridge University Press.

Academic Year 2020-21

Course unit

Courses INFORMATICS ENGINEERING (*)

(*) Optional course unit for this course

Faculty / School FACULTY OF SCIENCES AND TECHNOLOGY

Main Scientific Area CIÊNCIAS DE COMPUTAÇÃO

Acronym

Language of instruction Portuguese (PT) or English (ENG)

Teaching/Learning modality On-line

Coordinating teacher António Eduardo de Barros Ruano

Teaching staff	Type	Classes	Hours (*)
António Eduardo de Barros Ruano	PL; T	T1; PL1	28T; 28PL

* For classes taught jointly, it is only accounted the workload of one.

Contact hours

T	TP	PL	TC	S	E	OT	O	Total
28	0	28	0	0	0	0	0	156

T - Theoretical; TP - Theoretical and practical ; PL - Practical and laboratorial; TC - Field Work; S - Seminar; E - Training; OT - Tutorial; O - Other

Pre-requisites

no pre-requisites

Prior knowledge and skills

Neural Networks and Deep Learning

The students intended learning outcomes (knowledge, skills and competences)

The student, at the end of the course, should be able to:

- Process different types of data (high-dimensional, graphs, infinite, labeled)
- Use different models of computation (MapReduce, Streams and online algorithms)
- Solve real-word problems (Recommender Systems, Market Basket Analysis, Spam Detection, Duplicate Document detection) ´

Syllabus

1. Introduction - what is Big Data processing?
2. Hadoop/MapReduce Programming
3. Finding Similar Items
4. Mining Data Streams
5. Link Analysis
6. Frequent Itemsets
7. Clustering
8. Advertising on the Web
9. Recommendation Systems
10. Mining Social-Network Graphs
11. Dimensionality Reduction
12. Large-Scale Machine Learning

Teaching methodologies (including evaluation)

Classes will be distributed by Theoretical (T) and Laboratory Practice (PL) components. In lectures T the theory will be exposed, and examples of application will be presented. In PL classes students will implement programs in Hadoop, R or Python, using the available modules in each of these languages/environments.

In this discipline there will be a group mini-project, and an individual project, both with a weight of 50% of the final grade. If the final grade is equal to or higher than 10, and both components are graded with a mark higher than 8, the student passes.

Main Bibliography

On-line Lecture Notes

Marz, N. and J. Warren, Big Data: Principles and best practices of scalable realtime data systems. 2015: Manning Publications.

Hwang, K. and M. Chen, Big-Data Analytics for Cloud, IoT and Cognitive Computing. 2017.

Ratner, B., Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. 3 ed. 2017: CRC Press.

Lin, J., & Dyer, C. (2010). Data-Intensive Text Processing with MapReduce: Morgan and Claypool Publishers.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of Massive Datasets: Cambridge University Press.