Course guide
300271 - BIGDATA - Big Data & Data Mining

Last modified: 06/06/2024

Unit in charge: Castelldefels School of Telecommunications and Aerospace Engineering
Teaching unit: 701 - DAC - Department of Computer Architecture.

Degree: MASTER'S DEGREE IN APPLIED TELECOMMUNICATIONS AND ENGINEERING MANAGEMENT (MASTEAM) (Syllabus 2015). (Optional subject).

Academic year: 2024 **ECTS Credits:** 6.0 **Languages:** English

LECTURER

Coordinating lecturer:

Others:

PRIOR SKILLS

English, Programming, Probability.

REQUIREMENTS

English, Programming, Probability.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Generical:
03 DIS. (ENG) Diseñar aplicaciones de alto valor añadido basadas en las Tecnologías de la Información y las Comunicaciones (TIC), aplicadas a cualquier ámbito de la sociedad.
06 RES. (ENG) Resolver problemas y mejorar procesos en cualquier ámbito social a partir de la aplicación de las TIC, integrando conocimientos de diversos ámbitos y aplicando ingeniería de alto nivel tecnológico.

Transversal:
05 TEQ. TEAMWORK. Being able to work as a team player, either as a member or as a leader. Contributing to projects pragmatically and responsibly, by reaching commitments in accordance to the resources that are available.
03 TLG. THIRD LANGUAGE. Learning a third language, preferably English, to a degree of oral and written fluency that fits in with the future needs of the graduates of each course.

Basic:
CB6. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

TEACHING METHODOLOGY

The course is organized as a hands-on subject in which students work on projects related to the Big Data analytics. The main methodology is project based learning.

LEARNING OBJECTIVES OF THE SUBJECT

At the end of the course the student should be able to apply a number of data mining technologies over large data sets, extract useful information out of big data, program using the map-reduce paradigm and execute at large scale using cluster/cloud computers.

STUDY LOAD

Type	Hours	Percentage
Hours small group	54,0	36.00
Self study	96,0	64.00

Total learning time: 150 h

CONTENTS

T1

Description:
Introduction to Big Data: Presentation of the course, examples of usage of big data technologies, available resources and developing environments.

Related activities:
A1

Full-or-part-time: 10h
Laboratory classes: 4h
Self study : 6h

T2

Description:
Data sources, distributed file systems and databases, and data streaming: Technologies on Indexing, Memory, Streams, databases and evolution to big data. First examples on input sets.

Related activities:
A1+A2

Full-or-part-time: 25h
Laboratory classes: 15h
Self study : 10h

T3

Description:
Processing and Data mining: Basic foundations and applications of map-reduce programming, learning models (search, classification, regression, clustering, information extraction), Bayesian inference, logic of reasoning, uncertainties and forecasting.

Related activities:
A2

Full-or-part-time: 115h
Laboratory classes: 35h
Self study : 80h

ACTIVITIES

A1

Description:
Guided exercises: Install of the programming environment and big data tools (i.e. Apache tools), basic examples and programs: hello world, lists, dictionaries, etc. Set up of data and machine learning libraries.

Material:
Atenea

Delivery:
A1 in Atenea (30%)

Related competencies :
06 RES. (ENG) Resolver problemas y mejorar procesos en cualquier ámbito social a partir de la aplicación de las TIC, integrando conocimientos de diversos ámbitos y aplicando ingeniería de alto nivel tecnológico.

Full-or-part-time: 26h
Laboratory classes: 10h
Self study: 16h

A2

Description:
Project: Classify objects based on features, using a variety of methods. Use Decision Trees and Bayesian Networks to explain phenomenon. Predict indicators using regression techniques. Display and analyze groups in your data using dimensionality reduction. Pre-process, extract, and select the learning features. Select the best parameters for your models using model selection.

Material:
Atenea

Delivery:
A2 in Atenea (70%)

Related competencies :
03 DIS. (ENG) Diseñar aplicaciones de alto valor añadido basadas en las Tecnologías de la Información y las Comunicaciones (TIC), aplicadas a cualquier ámbito de la sociedad.
06 RES. (ENG) Resolver problemas y mejorar procesos en cualquier ámbito social a partir de la aplicación de las TIC, integrando conocimientos de diversos ámbitos y aplicando ingeniería de alto nivel tecnológico.
05 TEQ. TEAMWORK. Being able to work as a team player, either as a member or as a leader. Contributing to projects pragmatically and responsibly, by reaching commitments in accordance to the resources that are available.
03 TLG. THIRD LANGUAGE. Learning a third language, preferably English, to a degree of oral and written fluency that fits in with the future needs of the graduates of each course.
CB6. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

Full-or-part-time: 124h
Laboratory classes: 44h
Self study: 80h

GRADING SYSTEM

A1=30% + A2=70%

EXAMINATION RULES.

Students should attend with their own personal laptop. Assistance is mandatory for at least 80% of class time. Activities are done in group.

BIBLIOGRAPHY

Basic:
- Géron, Aurélien. Hands-on machine learning with scikit-learn & tensorflow : concepts, tools, and techniques to build intelligent systems [on line]. Sebastopol, CA: O'Reilly Media, Inc, [2017] [Consultation: 26/07/2022]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=4822582. ISBN 9781491962299.

Complementary:
- Macias Lloret, Mario; Gómez Mauro; Tous Liesa, Rubén; Torres, Jordi. Introducción a Apache Spark : para empezar a programar el big data. Barcelona: UOC, 2015. ISBN 9788491160373.
- Mohanty, Hrushikesha; Bhuyan, Prachet; Chenthati, Deepak. Big Data : A Primer [on line]. New Delhi: Springer India, 2015 [Consultation: 26/07/2022]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-81-322-2494-5. ISBN 9788132224945.
- Leskovec, Jure; Rajaraman, Anand; Ullman, Jeffrey D. Mining of massive datasets [on line]. 2nd ed. New York, N.Y. ; Cambridge: Cambridge University Press, 2014 [Consultation: 26/07/2022]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=807230. ISBN 9781107077232.
- Garreta, Raúl; Moncecchi, Guillermo. Learning scikit-learn : machine learning in Python. Birmingham: Packt Publishing, 2013. ISBN 9781783281930.

RESOURCES

Other resources:
Atenea