Universitat de Lleida

GUIA DOCENT

PROJECTE BIG DATA

Coordinació: GARCIA GONZALEZ, ROBERTO

Any acadèmic 2019-20

PROJECTE BIG DATA 2019-20

Informació general de l'assignatura

Denominació	PROJECTE BIG DATA
Codi	103090
Semestre d'impartició	1R Q(SEMESTRE) AVALUACIÓ CONTINUADA
Caràcter	

Grau/Màster	Curs	Caràcter	Modalitat
Màster Universitari en Enginyeria Informàtica (M 2015)	2	OPTATIVA	Presencial

Nombre de crèdits assignatura (ECTS)	6
Tipus d'activitat, crèdits i grups	

Tipus d'activitat	PRALAB	TEORIA
Nombre de crèdits	3	3
Nombre de grups	1	1

Coordinació	GARCIA GONZALEZ, ROBERTO
Departament/s	INFORMÀTICA I ENGINYERIA INDUSTRIAL
Distribució càrrega docent entre la classe presencial i el treball autònom de l'estudiant	30% temps sessions presencials (3 hores/setmana) i el 70% restant es basa en treball autònom de l'estudiant.
Informació important sobre tractament de dades	Consulteu aquest enllaç per a més informació.
Idioma/es d'impartició	Anglès

PROJECTE BIG DATA 2019-20

Professor/a (s/es)	Adreça electrònica professor/a (s/es)	Crèdits impartits pel professorat	Horari de tutoria/lloc
COMAS RODRIGUEZ, CARLOS	carles.comas@udl.cat	3	
GARCIA GONZALEZ, ROBERTO	roberto.garcia@udl.cat	0	
VIRGILI GOMÀ, JORDI	jordi.virgili@udl.cat	3	

Informació complementària de l'assignatura

Per a un millor aprofitament d'aquesta assignatura, es recomana que l'estudiant tingui coneixements de programació en Python i que hagi completat l'assignatura Massive Data Processing.

Objectius acadèmics de l'assignatura

1. Comprendre i aplicar tècniques estadístiques de mineria de dades.
2. Utilitzar adequadament els paquets estadístics per a l'anàlisi de dades.
3. Comprendre i aplicar tècniques estadístiques relacionades amb el funcionament de grans volums de dades.
4. Utilitzar adequadament els paquets estadístics per a aquest tipus d'explotació.
5. Proposar una adequada visualització de la informació analitzada que facilita la comprensió.
6. Conèixer les aplicacions comunes en l'àmbit de grans volums de dades i ser capaç de desenvolupar solucions per a aquests problemes.
7. Sistemàticament desplegar el conjunt de tècniques comunes per resoldre problemes grans de dades.
8. Comunicar eficaçment els resultats del projecte als socis tècnics i clients.

Competències

Competències Generals

- **CG4:** Capacitat per al modelatge matemàtic, càlcul i simulació en centres tecnològics i d'enginyeria d'empresa, particularment en tasques d'investigació, desenvolupament i innovació en tots els àmbits relacionats amb l'Enginyeria en Informàtica.
- **CG8:** Capacitat per a l'aplicació dels coneixements adquirits i de resoldre problemes en entorns nous o poc coneguts dins de contextos més amplis i mulitidisciplinares, sent capaços d'integrar aquests coneixements.

Competències Estratègiques de la UdL

- **UdL2**: Domini d'una llengua estrangera.

Competències transversals

- **EPS1:** Capacitat de planificació i organització del treball personal.
- **EPS3:** Capacitat de transmetre informació, idees, problemes i solucions a un públic tant especialitzat com no especialitzat.
- **EPS4:** Capacitat de concebre, dissenyar i implementar projectes i/o aportar solucions noves, utilitzant eines

pròpies de l'enginyeria.
- **EPS5:** Tenir motivació per la qualitat i la millora contínua.

Competències Bàsiques

- **CB3**: Ser capaç d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, sent incompleta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.
- **CB4**: Saber comunicar les conclusions -i els coneixements i raons últimes que les sustenten- a públics especialitzats i no especialitzats d'una manera clara i sense ambigüitats.

Competències Específiques

- **CE1**: Capacitat per a la integració de tecnologies, aplicacions, serveis i sistemes propis de l'Enginyeria Informàtica, amb caràcter generalista, i en contextos més amplis i multidisciplinaris.
- **CE2**: Capacitat per a la planificació estratègica, elaboració, direcció, coordinació, i gestió tècnica i econòmica en els àmbits de l'enginyeria informàtica relacionats, entre altres, amb: sistemes, aplicacions, serveis, xarxes, infraestructures o instal·lacions informàtiques i centres o factories de desenvolupament de programari, respectant l'adequat compliment dels criteris de qualitat i mediambientals i en entorns de treball multidisciplinaris.
- **CE4**: Capacitat per modelar, dissenyar, definir l'arquitectura, implantar, gestionar, operar, administrar i mantenir aplicacions, xarxes, sistemes, serveis i continguts informàtics.
- **CE5**: Capacitat de comprendre i saber aplicar el funcionament i organització d'Internet, les tecnologies i protocols de xarxes de nova generació, els models de components, programari intermediari i serveis.
- **CE7:** Capacitat per dissenyar, desenvolupar, gestionar i avaluar mecanismes de certificació i garantia de seguretat en el tractament i accés a la informació en un sistema de processament local o distribuït.

Continguts fonamentals de l'assignatura

1. Introducció als algoritmes PCA (Principal Component Analysis) i EM (Expectation-Maximization)
2. Principal Component Analysis
    1. Matrius de dades i espais associats
    2. Principal Component Analysis
    3. Interpretació i qualitat de resultats PCA
3. Algoritme Expectation-Maximization
    1. Maximum Likehood Estimation (MLE)
    2. Algoritme Expectation-Maximization (EM)
    3. EM per Missing Data
    4. Gaussian Mixture Model (EM Clustering)
4. Business Intelligence
    1. Recolçar decissions empresarils en Big Data
    2. Smart Data
    3. Visualització amb matplotlib
    4. Databricks, Google Cloud,...
5. Exploració interactiva de Big Data
    1. Apache Spark per consultar i explorar interactivament grans volums de dades heterogènies
    2. SparkSQL
    3. Spark R i Exploratory Data Analysis

Eixos metodològics de l'assignatura

Tots els cursos del bloc d'anàlisi del Big Dades (inclòs aquest), seran avaluats per un únic, comú, projecte que involucra a totes les assignatures (recopilació de dades, processament, aprenentatge, estadística, visualització, etc.).

Els estudiants treballaran en aquest projecte des del principi fins als últims cursos. Durant els cursos regulars, s'introduiran diferents temes, mostrant la seva relació amb el projecte comú i com tots els temes encaixen entre si

per crear una tasca o un projecte complex del món real.

Els tres cursos de formació de Big Data Analytics utilitzaran la mateixa configuració de base tecnològica:

- Python com a llenguatge de programació de base.
- Hadoop / Spark (amb Java, si és necessari)
- Si bé durant els cursos s'introduiran altres paquets tecnològics:
    - Scala
    - NodeJS
    - Etc.

Pla de desenvolupament de l'assignatura

Setmana	Descripció	Activitats Presencials	Treball Autònom Alumne
1	Introducció a CPA i algoritme EM	Presentació assignatura Classes magistrals i participatives	Estudi i resolució d'exercicis
2	Data Matrices i Associated Spaces	Classes magistrals i participatives	Estudi i resolució d'exercicis
3	PCA	Classes magistrals i participatives	Estudi i resolució d'exercicis
4	Interpertació i qualitat de resultats de PCA	Classes magistrals i participatives	Estudi i resolució d'exercicis
5	Maximum Likehood Estimation (MLE)	Classes magistrals i participatives	Estudi i resolució d'exercicis
6	Expectation-Maximization (EM) Algorithm	Classes magistrals i participatives	Estudi i resolució d'exercicis
7	EM per Missing Data i Gaussian Mixture Presentacions Orals	Classes magistrals i participatives Presentacions CPA i EM	Estudi i resolució d'exercicis Desenvolupament de projecte
8	Apache Spark Intro i Demos Revisió Big Data Project	Classes magistrals i participatives Presentacions Projectes	Estudi de casos i desenvolupament de projecte
9	Big Data Exploration	Classes magistrals i participatives	Desenvolupament de projecte
10	Big Data Exploration	Classes magistrals i participatives	Desenvolupament de projecte
11	Big Data Exploration	Classes magistrals i participatives	Desenvolupament de projecte
12	Big Data Exploration	Classes magistrals i participatives	Desenvolupament de projecte
13	Business Intelligence	Classes magistrals i participatives	Desenvolupament de projecte
14	Business Intelligence	Classes magistrals i participatives	Desenvolupament de projecte
15	Business Intelligence	Presentacions Projectes	Desenvolupament de projecte

Sistema d'avaluació

L'avaluació d'aquesta assignatura es basa en l'avaluació contínua.

ID	Activitats avaluació	%	Dates	Obligatòria	I/G (1)
OP1	Presentació Oral Resolució de problema pràctic (CPA)	25%	Setmana 9	Sí	Grup
OP2	Presentació Oral Resolució de problema pràctic (EM)	25%	Setmana 9	Sí	Grup
PP	Documentació Escrita i Presentació Oral Planificació Projecte	15%	Setmana 10	Sí	Grup

ID	Activitats avaluació	%	Dates	Obligatòria	I/G (1)
PD	Documentació Escrita Entregable Projecte	15%	Setmana 15	Sí	Grup
OP3	Presentació Oral Projecte	20%	Setmana 15	Sí	Grup

(1) Individual / Grup

La nota final es calcularà en base a la següent fórmula:

Nota Final = 0,25*OP1 + 0,25*OP2 + 0,15*PP + 0,15*PD + 0,2*OP3

Bibliografia i recursos d'informació

[Kar15] Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, "Learning Spark: Lightning-Fast Big Data Analysis", O'Reilly, 2015

[Ryz15] Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills, "Advanced Analytics with Spark: Patterns for Learning from Data at Scale", O'Reilly, 2015

[Bae14] Bart Baesens, "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications"