UNIVERSITÀ DEGLI STUDI DI PALERMO

DIPARTIMENTO	Ingegneria
ANNO ACCADEMICO OFFERTA	2019/2020
ANNO ACCADEMICO EROGAZIONE	2020/2021
CORSO DILAUREA MAGISTRALE	INGEGNERIA INFORMATICA
INSEGNAMENTO	BIG DATA
TIPO DI ATTIVITA'	B
AMBITO	50369-Ingegneria informatica
CODICE INSEGNAMENTO	19580
SETTORI SCIENTIFICO-DISCIPLINARI	ING-INF/05
DOCENTE RESPONSABILE	PIRRONE ROBERTO Professore Associato Univ. di PALERMO
ALTRI DOCENTI	
CFU	12
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	204
NUMERO DI ORE RISERVATE ALLA DIDATTICA ASSISTITA	96
PROPEDEUTICITA'	
MUTUAZIONI	
ANNO DI CORSO	2
PERIODO DELLE LEZIONI	1° semestre
MODALITA' DI FREQUENZA	Facoltativa
TIPO DI VALUTAZIONE	Voto in trentesimi
ORARIO DI RICEVIMENTO DEGLI STUDENTI	**PIRRONE ROBERTO** Mercoledì 11:30 13:00 Studio del docente, Edificio 6, terzo piano, stanza 3025

DOCENTE: Prof. ROBERTO PIRRONE

PREREQUISITI	
RISULTATI DI APPRENDIMENTO ATTESI	Conoscenza e capacita' di comprensione Lo studente, al termine del corso, avra' acquisito conoscenze e metodologie per affrontare le problematiche legate sia all'analisi dei piu' diffusi tipi di dati sia all'utilizzo di architetture software per la gestione dei Big Data. Lo studente conoscera' in maniera adeguata le differenze tra i diversi algoritmi di analisi in relazione alla tipologia dei dati, conoscera' le tecniche di pre-processing piu' adatte e come definire l'architettura Big Data piu' efficiente per condurre le proprie analisi. Per il raggiungimento di quest'obiettivo il corso comprende un ciclo di lezioni frontali sugli argomenti della disciplina. Per la verifica di quest'obiettivo l'esame comprende la presentazione orale dei casi di studio e la discussione orale. Capacita' di applicare conoscenza e comprensione Lo studente avra' acquisito conoscenze e metodologie per analizzare e risolvere problemi tipici legati alla l'implementazione di pipeline complete di analisi dei dati sia per dataset classici sia per Big Data. Egli avra' profonda conoscenza del linguaggio di programmazione Python e delle principali librerie per l'analisi e la visualizzazione dei dati quali Numpy, SciPy, Scikit-learn, Matplotlib, Pandas, Tensorflow e Keras. Inoltre lo studente avra' sufficiente conoscenza dei database noSQL quale Apache Cassandra e del framework per Big Data Apache Hadoop con il suo ecosistema, mentre acquisira' profonda conoscenza del farmework Apache Spark e delle sue librerie di analisi dei dati e di interazione con i database nella loro interfaccia Python. Per il raggiungimento di quest'obiettivo il corso comprende: esercitazioni teoriche e di gruppo per sviluppo di applicazioni web complesse con tecnologia Javascript e PHP. Per la verifica di quest'obiettivo l'esame comprende la presentazione degli elaborati software preparati durante le esercitazioni di gruppo. Autonomia di giudizio Lo studente sara' in grado di svolgere un'analisi comparativa delle caratteristiche di differenti ambienti e/o infrastrutture di analisi di Big Data in relazione alla soluzione di problemi specifici. Egli sara' in grado di affrontare a livello operativo problemi non strutturati e prendere decisioni in regime d'incertezza. Attraverso l'approccio metodologico acquisito durante il corso, egli potra' condurre lo sviluppo di nuove problematiche applicative nell'ambito dei Big Data e della data analysis in generale. Per il raggiungimento di quest'obiettivo il corso comprende: la presentazione e discussione in aula di progetti e implementazioni legati alle esercitazioni pratiche di gruppo. Per la verifica di quest'obiettivo l'esame comprende la discussione sui casi di studio presentati e su possibili varianti proposte dal docente. Abilita' comunicative Lo studente sara' in grado di comunicare con competenza e proprieta' di linguaggio problematiche complesse di data analysis e Big Data. Per il raggiungimento di quest'obiettivo il corso comprende: esercitazioni di gruppo per sviluppo di un'intera pipeline di analisi di dati con tecnologie per i Big Data su un caso di studio proposto dal docente nonche' la presentazione e discussione in aula di progetti e implementazioni. Per la verifica di quest'obiettivo l'esame comprende la discussione sui casi di studio presentati e su possibili varianti proposte dal docente. Capacita' d'apprendimento Lo studente sara' in grado di affrontare in autonomia qualsiasi problematica concernente lo sviluppo di pipeline complete per analisi di Big Data. Sara' in grado di approfondire tematiche complesse legate all'analisi di prestazioni di framework diversi cogliendone i punti di forza e di debolezza. Per il raggiungimento di quest'obiettivo il corso comprende: esercitazioni di gruppo per sviluppo di un'intera pipeline di analisi di dati con tecnologie per i Big Data su un caso di studio proposto dal docente nonche' la presentazione e discussione in aula di progetti e implementazioni. Per la verifica di quest'obiettivo l'esame comprende la discussione sui casi di studio presentati e su possibili varianti proposte dal docente.
VALUTAZIONE DELL'APPRENDIMENTO	L'esame finale consta di due parti: la presentazione delle applicazioni sviluppate come caso di studio proposto dal docente e la prova orale. La presentazione dei casi di studio verra' valutata secondo i seguenti aspetti del codice prodotto: •Completezza dell'analisi condotta •Trattamento dei dati •Feature selection •Originalita

	•Capacita' di integrazione di codice gia' noto dalle esercitazioni teoriche •Performance ottenuta. Inoltre, l'esposizione orale dei casi di studio da parte dei singoli studenti verra' valutata rispetto ai seguenti aspetti: •Grado di comprensione mostrato in relazione al programma teorico svolto •Proprieta' del linguaggio utilizzato •capacita' di approfondimento del tema assegnato mediante letture autonome. La presentazione del caso di studio si intende superata se la valutazione e' di almeno 18/30 ed e' preclusiva dell'accesso alla prova orale. Il colloquio orale tende a verificare le conoscenze dei temi esposti nel programma teorico svolto. Il voto finale risulta da una media delle valutazioni riportate nelle due parti dell'esame poiche' queste coprono aspetti diversi ed egualmente importanti della preparazione dello studente.
OBIETTIVI FORMATIVI	Il corso di "Big Data" fornisce agli studenti una conoscenza approfondita delle architetture software per i Big Data nonche' dei principali algoritmi di analisi dei dati e delle tecniche di preprocessing di tali dati, al fine di sviluppare autonomamente intere pipeline di analisi per dei casi di studio reali. Il corso consente di acquisire 12 CFU e consta di una serie di lezioni ed esercitazioni teoriche e la costituzione di gruppi di lavoro per l'analisi di casi di studio proposti dal docente attraverso lo sviluppo di un'intera pipeline di analisi di dati con tecnologie per i Big Data. Il risultato dell'attivita' dei gruppi di lavoro viene poi discusso coralmente in aula. Il ciclo di lezioni teoriche presenta dapprima un'introduzione al processo di analisi dei dati nel suo complesso e fornisce alcuni cenni sui datawarehouse. Successivamente si affrontano le tecniche di prerocessing dei dati quali la riduzione di dimensionalita' e la gestione di dati mancanti e si introducono alcune misure di similarita' piu' diffusamente usate nel campo della data analysis. Si passa ad affrontare il clustering e i classificatori nonche' le reti neurali e il deep learning. Segue la presentazione di tipiche pipeline di processo per dati di varia natura: testi sequenze temporali e discrete, grafi, dati web. L'ultima parte del corso e' dedicata propriamente alle architetture software per i Big Data: si affronteranno i database noSQL, l'algoritmo MapReduce e Apache Hadoop e si trattera' estensivamente il framework Apache Spark. Le esercitazioni teoriche coprono la configurazione degli ambienti di sviluppo con cui si operera' durante il corso e l'illustrazione dei temi affrontati nel corso teorico attraverso esempi svolti. Infine i gruppi di lavoro sono mirati allo sviluppo di pipeline complesse di analisi di Big Data su dei casi di studio reali.
ORGANIZZAZIONE DELLA DIDATTICA	Lezioni frontali; Esercitazioni teoriche; Esercitazioni di gruppo per lo sviluppo di pipeline di analisi dei dati con tecnologie Big Data.
TESTI CONSIGLIATI	Data Mining: The Textbook, 2015, Charu C. Aggarwal, Springer-Verlag New York, ISBN 978-3319141411, prezzo orientativo € 70,00 Deep Learning, (2016), di Ian Goodfellow, Yoshua Bengio, Aaron Courville, MIT Press, ISBN 978-0262035613, prezzo orientativo €65,00 Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition, (2017) Sebastian Raschka, Vahid Mirjalili, Packt Publishing, ISBN 978-1787125933, prezzo orientativo € 35,00 Spark: The Definitive Guide: Big Data Processing Made Simple, 2018, di Bill Chambers e Matei Zaharia, Oreilly & Associates Inc, ISBN 978-1491912218, prezzo orientativo € 45,00. Materiale didattico in forma elettronica disponibile sul portale di Ateneo Siti web con manuali di riferimento per le esercitazioni ed i testi: https://link.springer.com/book/10.1007%2F978-3-319-14142-8 http://www.deeplearningbook.org/ https://github.com/PacktPublishing/Python-Machine-Learning-Second-Edition https://github.com/databricks/Spark-The-Definitive-Guide

PROGRAMMA

ORE	Lezioni
2	Introduzione al Corso. Il processo di analisi dei dati: raccolta dei dati, pre-processing, applicazione delle tecniche di analisi ed estrazione della conoscenza.

PROGRAMMA

ORE	Lezioni
3	Preparazione dei dati: tipi di dati, data cleaning, gestione dei dati mancanti, campionamento.
3	Riduzione della dimensionalita: Principal Component Analysis, Singular Value Decomposition, Trasformazioni Wavelet, Multi Dimensional Scaling, Embedding di grafi.
2	Distanze e similarita' per i diversi tipi di dati: dati quantitativi, dati categoriali, dati testuali, sequenze temporali, grafi.
2	Data cubes, Cenni sulla tecnologia OLAP e creazione di un datawarehouse.
3	Mining di pattern ricorrenti: algoritmo Apriori, misure statistiche di correlazione.
5	Clustering: k-means e simili, clustering gerarchico, clustering density based e a griglia, clustering basato su grafi, clustering di dati ad elevata dimensionalita, validazione del clustering, analisi degli outlier.
5	Classificatori: feature selection, decision tree e classificatori a regole, Naive Bayes, regressione logistica, Support Vector Machines, Nearest Neighbor, valutazione dei classificatori.
5	Classificatori, concetti avanzati: Multi-class e rare class learning, regressione su dati numerici, semi-supervised learning, metodi di ensemble.
15	Deep Learning: reti feed forward profonde: funzioni di attivazione, livelli nascosti, architettura, backrpopagation e altri metodi di apprendimento. Ottimizzazione per reti profonde: discesa stocastica del gradiente; strategie di inizializzazione dei parametri, algoritmi adattivi; strategie di ottimizzazione e meta-algoritmi. Reti convolutive: l'operazione di convoluzione, pooling, varianti della funzione di convoluzione, tipo di dati, algoritmi di convoluzione efficienti, caratteristiche random o non supervisionate. Reti ricorrenti e ricorsive: grafi computazionali, reti neurali ricorrenti, reti ricorrenti bidirezionali, architetture encoder-decoder sequence-to-sequence, reti ricorrenti profonde, reti ricorsive, reti ricorrenti Long-Short-Term Memory. Autocodificatori e reti generative: macchine di Boltzmann, macchine di Boltzman ristrette, reti deep belief, macchine di Boltzmann profonde, macchine di Boltzmann convolutive.
2	Analisi di dati testuali: preprocessing dei documenti, LSA, Naive Bayes, SVM.
2	Analisi di serie temporali: preprocessing, trasformate, modelli autoregressivi, clustering.
2	Discrete sequence analysis: Markov process, Hidden Markov Models, Kernel SVM. Analisi di sequenze discrete: processi markoviani, Hidden Markov Models, Kernel SVM.
2	Analisi di grafi: distanze tra grafi, matching tra grafi, clustering e classificazione.
2	Analisi di dati web: algoritmo PageRank, recommender systems, web usage analysis, social network analysis.
6	Architetture software per i Big Data: database noSQL, Apache Cassandra, MongoDB, Neo4j.
6	Architetture software per i Big Data: l’algoritmo MapReduce, Apache Hadoop e il suo ecosistema (Pig, Hive, HDFS).
6	Architetture software per i Big Data: Spark e le sue librerie.

ORE	Esercitazioni
2	Installazione e configurazione dell’ambiente Python e delle principali librerie di analisi dei dati
4	Sviluppo di una pipeline di analisi di dati testuali e da social media
3	Sviluppo di una pipeline di analisi di dati biologici sequenziali
3	Sviluppo di una pipeline di analisi di dati farmacologici (grafi 2D/3D)
3	Sviluppo di una pipeline di analisi di segnali biomedici
3	Sviluppo di una pipeline di analisi di immagini mediche attraverso reti convoluzionali.
2	Sviluppo di una pipeline di analisi del linguaggio naturale attraverso reti ricorrenti LSTM
2	Installazione e configurazione di Spark e delle sue librerie
3	Sviluppo di una pipeline di analisi dei dati su Spark
3	Sviluppo di una pipeline di deep learning su Spark

ORE	Altro
6	Sviluppo di un’intera pipeline di analisi di dati con tecnologie per i Big Data su un caso di studio proposto dal docente.