UNIVERSITÀ degli STUDI di CATANIA

DIPARTIMENTO DI MATEMATICA E INFORMATICA
Corso di laurea magistrale in Informatica
Anno accademico 2016/2017 - 1° anno

BIG DATA

INF/01 - 6 CFU - 2° semestre

Docente titolare dell'insegnamento

ALFREDO PULVIRENTI
Email: apulvirenti@dmi.unict.it
Edificio / Indirizzo: Stanza 35, Terzo Blocco Dipartimento di Matematica e Informatica.
Telefono: 095-7383087
Orario ricevimento: Martedi' 10-11.

OBIETTIVI FORMATIVI

Obiettivi formativi generali dell'insegnamento in termini di risultati di apprendimento attesi.

1. **Conoscenza e capacità di comprensione (knowledge and understanding):** Il corso mira a formare le conoscenze e le competenze di base ed avanzate per l'analisi per la rappresentazione, l'organizzazione e l'analisi di grandi quantità di dati.
2. **Capacità di applicare conoscenza e comprensione (applying knowledge and understanding):** lo studente acquisirà conoscenze riguardo ai modelli e gli algoritmi per l'analisi dei dati quali: mining ad alto supporto, sistemi di raccomandazione, ricerca di similarità ad alte dimensioni, map-reduce e spark, complex network analysis, text mining e sistemi per il tagging di documenti.
3. **Autonomia di giudizio (making judgements):** Attraverso esempi concreti e casi di studio, lo studente sarà in grado di elaborare autonomamente soluzioni a determinati problemi legati ai big data.
4. **Abilità comunicative (communication skills):** lo studente acquisirà le necessarie abilità comunicative e di appropriatezza espressiva nell'impiego del linguaggio tecnico nell'ambito generale dei big data.
5. **Capacità di apprendimento (learning skills):** il corso si propone, come obiettivo, di fornire allo studente le necessarie metodologie teoriche e pratiche per poter affrontare e risolvere autonomamente nuove problematiche che dovessero sorgere durante una attività lavorativa. A tale scopo diversi argomenti saranno trattati a lezione coinvolgendo lo studente nella ricerca di possibili soluzioni a problemi reali, utilizzando benchmark disponibili in letteratura.

PREREQUISITI RICHIESTI

Programmazione, strutture dati, algoritmi su grafi, concetti di base di probabilità e statistica.

FREQUENZA LEZIONI

Le risorse principali messe a disposizione dello studente sono le **lezioni frontali**, la cui frequenza è **fortemente consigliata**.

Per seguire meglio le lezioni, vengono messe a disposizione le **slide** utilizzate per il corso. Le slide non costituiscono un mezzo di studio: forniscono un dettaglio puntuale sugli argomenti trattati a lezione.

CONTENUTI DEL CORSO

- Mining alto supporto:
  - apriori
  - sue estensioni: PCY, estensioni iceberg
  - generazione degli insiemi frequenti senza il calcolo dei candidati
- Sistemi di raccomandazione:
  - collaborative filtering
  - NBI
  - Modelli ibridi
- Map-Reduce:
  - Concetti, motivazioni e algoritmi:
  - conteggio parole documenti, prodotto vettore/matrice; Prodotto matrice/matrice; Join; Group By
  - hadoop
  - Beyond map-reduce
- Ricerca di similarità su alte dimensioni:
  - Shingling
  - Min-Hashing
  - LSH
  - Min-LSH
- Dimensionality reduction:
  - PCA
  - SVD
  - CUR
  - Applicazione LSI
  - Teorema di Johnson-Lindenstrauss
- Link Analysis:
  - PageRank
  - Link spam
  - Hub-Authorities
  - Applicazioni su Map-Reduce
- Web Advertising:
  - Algoritmi online
  - Adword e sue implementazioni
- Graph mining e network analysis:
  - Conteggio triangoli
  - subgraph matching e motif finding

    - community detection: overlapping communities
    - Network alignment
- Text mining
    - Cenni sul text mining: TF.IDF, Bag-Of-Word, Entity annotation
    - Collegamento con dimensionality reduction
- Analisi dei dati in R attraverso la programmazione funzionale

---

TESTI DI RIFERIMENTO

Mining of Massive Datasets

Jure Leskovec, Anand Rajaraman, Jeff Ullman

http://www.mmds.org

ALTRO MATERIALE DIDATTICO

Il materiale didattico sarà pubblicato su www.studium.unict.it

---

PROGRAMMAZIONE DEL CORSO

	*	**Argomenti**	**Riferimenti testi**
1	*	Introduzione, Map Reduce, Spark	Capitoli 1 e 2 + materiale didattico integrativo
2	*	Mining di insiemi frequenti	Capitolo 6 + materiale didattico integrativo
3	*	Similarità ad alte dimensioni. Locality sensitive Hashing (LSH).	Capitolo 3 + materiale didattico integrativo
4	*	Attività pratica su LSH e sue applicazioni	
5	*	Dimensionality reduction. PCA, SVD, CUR, NNMF	Capitolo 11 + materiale didattico integrativo
6	*	Attività pratica su dimensionality reduction.	
7	*	Sistemi di raccomandazione. Latent Semantic Indexing, Collaborative filtering e Network based inference,	Capitolo 9 + materiale didattico integrativo
8	*	Attività pratica su sistemi di raccomandazione.	
9	*	Link Analysis: PageRank Link spam Hub-Authorities Applicazioni su Map-Reduce	Capitolo 5 + materiale didattico integrativo

10	*	Analisi di Grafi di grandi dimensioni. Conteggio triangoli subgraph matching e motif finding, community detection: overlapping communities Network alignment	Capitolo 10 + materiale didattico integrativo
11	*	Attività pratica su motif finding su grafi di grandi dimensioni. Applicazioni in Finanza.	
12	*	Web Advertising: Algoritmi online Adword e sue implementazioni	Capitolo 8 è materiale didattico integrativo
13	*	Text mining. TF.IDF, Entity annotation	Materiale didattico integrativo
14	*	Attività pratica su text mining e sistemi di raccomandazione per analisi di banche dati citazioni: arxiv, pubmed	

* Conoscenze minime irrinunciabili per il superamento dell'esame.

N.B. La conoscenza degli argomenti contrassegnati con l'asterisco è condizione necessaria ma non sufficiente per il superamento dell'esame. Rispondere in maniera sufficiente o anche più che sufficiente alle domande su tali argomenti non assicura, pertanto, il superamento dell'esame.

VERIFICA DELL'APPRENDIMENTO

MODALITÀ DI VERIFICA DELL'APPRENDIMENTO

L'esame finale consiste in **una prova scritta** ed un **colloquio orale** nel quale viene discusso un progetto.

La prova scritta è costituita da esercizi e domande di teoria.

Chi non supera la prova scritta, non può sostenere l'orale. La prova scritta può essere visionata prima delle prove orali.

Salvo diversa comunicazione:

- l'esame scritto si svolge alle **ore 9:00**

Note:

- È **vietato** l'uso di qualsiasi strumento hardware (calcolatrici, tablet, smartphone, cellulari, auricolari BT etc.), di libri o documenti personali durante gli esami (scritti).
- Per sostenere gli esami è **obbligatorio prenotarsi** utilizzando l'apposito modulo del portale CEA.
- Non sono ammesse prenotazioni tardive tramite email. In mancanza di prenotazione, l'esame non può essere verbalizzato.

PROVE IN ITINERE

Non previste.

PROVE DI FINE CORSO

non previste.

ESEMPI DI DOMANDE E/O ESERCIZI FREQUENTI

Esempi saranno pubblicati sul portale www.studium.unict.it