UNIVERSITÀ degli STUDI di CATANIA

DIPARTIMENTO DI MATEMATICA E INFORMATICA
Corso di laurea magistrale in Informatica
Anno accademico 2018/2019 - 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

Il corso introduce le principali tecniche di data mining. Il focus è sul mining di grandi quantità di dati, tali da non entrare in memoria principale. I diversi esempi presentati durante il corso riguarderanno il web, le social network e i dati Next Generation Sequencing prodotti in ambito biomedico. Inoltre il corso affronta la tematica anche dal punto di vista degli algoritmi, enfatizzando la differenza dal machine-learning. Tra gli argomenti affrontati troviamo, in primo luogo tool quali, map-reduce, per lavorare in ambito distribuito con grandi quantità di dati. Tale argomento farà da denominatore comune in tutte le problematiche di mining presentate. Successivamente viene affrontato il problema della ricerca di similarità e dell'uso delle tecniche di hashing per grandi volumi di dati. Si affronta anche il problema classico del mining ad alto supporto descrivendo l'algoritmo apriori e le sue varianti. Saranno quindi introdotti i sistemi di raccomandazione. In tale contesto si affronterà pure il problema dei dati ad alta dimensionalità e le tecniche riduzione della dimensionalità quali, SVD, CUR, NNMF. Il corso introdurrà quindi le tematiche principali nell'analisi delle network. Verranno introdotte le misure di centralità per le network, con particolare riferimento al page-rank e alle sue varianti. Sarà introdotto il concetto di modello nullo di network in grado di conservare le caratteristiche della rete quali, distribuzione dei degree e coefficiente di clustering. Tra i modelli presentati troveremo: Erdos-Renyi, Chung-Lu, Preferential Attachment. Sarà affrontato il problema del clustering attraverso l'uso delle tecniche basate su modularità e clustering spettrale.

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.

MODALITÀ DI SVOLGIMENTO DELL'INSEGNAMENTO

lezioni frontali e laboratorio

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:
    - Modelli di reti Random
    - 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

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	

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.

ESEMPI DI DOMANDE E/O ESERCIZI FREQUENTI

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