UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Text Mining and Search

2526-2-FDS01Q013

Obiettivi

Conoscenza e comprensione

L'obiettivo dell'insegnamento è introdurre i concetti fondamentali dell'elaborazione del linguaggio naturale (*Natural Language Processing*, NLP) e delle tecniche di *Text Mining*, fornendo una base solida per l'analisi automatica dei testi. Dopo una panoramica sulle metodologie di base, l'insegnamento affronterà alcuni *task* specifici, tra cui la classificazione e il *clustering* di documenti, il *topic modeling*, la generazione automatica di riassunti (*text summarization*) e la ricerca di informazioni testuali.

Capacità di applicare conoscenza e comprensione

Durante le lezioni e le attività di laboratorio vengono promosse e valutate le capacità degli studenti di applicare le conoscenze acquisite sugli argomenti trattati nell'insegnamento.

Autonomia di giudizio

L'insegnamento mira a sviluppare l'autonomia di giudizio e la capacità di analisi critica rispetto alle principali sfide legate all'elaborazione del linguaggio naturale e al *Text Mining*, nonché ai principali *task* correlati. Tali competenze saranno stimolate anche attraverso discussioni in aula e attività di laboratorio.

Abilità comunicative

Sviluppo della capacità di comunicare in modo chiaro, consapevole e privo di ambiguità contenuti tecnici, idee, problemi e relative soluzioni a interlocutori diversi. Tali abilità saranno promosse durante l'insegnamento e valutate in sede d'esame.

Capacità di apprendimento

L'insegnamento è concepito per fornire sia conoscenze teoriche sia competenze pratiche, costituendo un solido

punto di partenza anche per eventuali approfondimenti individuali sui principi di rappresentazione, analisi e ricerca testuale.

Contenuti sintetici

- L'insegnamento fornirà una definizione introduttiva di *Text Mining* e *Natural Language Processing* (NLP), evidenziando le principali differenze tra *Data Mining* e *Text Mining*.
- Verranno presentate le principali tecniche di pre-processing testuale e affrontati i problemi legati all'indicizzazione dei testi e alla loro rappresentazione formale.
- Saranno quindi introdotte alcune applicazioni fondamentali del *Text Mining*, tra cui la classificazione e il *clustering* di documenti, il *topic modeling*, il riassunto automatico di testi e il reperimento delle informazioni testuali.
- Verranno inoltre presentati alcuni strumenti open source utili per lo sviluppo di applicazioni di *Text Mining*.

Programma esteso

1. Definizione di *Natural Language Processing* (NLP), *Text Mining* e principali differenze tra *Text Mining* e *Data Mining*.
2. Breve introduzione ad alcune applicazioni del *Text Mining*.
3. Tecniche di pre-processing, indicizzazione e rappresentazione formale dei testi (*Bag-of-Words*, *Word Embedding*, introduzione alle tecniche di *Contextualized Word Embedding*).
4. Classificazione e *clustering* di testi.
5. *Topic modeling*.
6. Riassunto automatico di testi.
7. Introduzione ai motori di ricerca testuali.
8. Strumenti "open source" per il *Text Mining* e la ricerca di informazioni online.

Prerequisiti

Conoscenze di base di statistica e di programmazione (preferibilmente in Python).

Modalità didattica

- L'insegnamento è costituito da 46 ore, di cui 9 di laboratorio.
- L'insegnamento è tenuto in lingua inglese.
- Le lezioni sono da 2 ore (una sola da 3 ore) e vengono svolte sia in modalità erogativa (specie nella parte iniziale della lezione) sia in modalità interattiva (durante la lezione) per il coinvolgimento attivo degli studenti.
- Nelle attività di laboratorio viene spiegato e sperimentato l'utilizzo di software “open source”.
- Potranno essere previsti seminari tenuti da esperti a livello nazionale ed internazionale.
- Alcune lezioni potranno essere svolte in modalità erogativa-interattiva da remoto.

Materiale didattico

- Berry, M. W., & Kogan, J. (Eds.). (2010). Text mining: applications and theory. John Wiley & Sons.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
- Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. Fundamentals of artificial intelligence, 603-649.

Altri testi specifici su Text Mining accesibili online verranno indicati durante l'insegnamento

Periodo di erogazione dell'insegnamento

Primo semestre.

Modalità di verifica del profitto e valutazione

Prova scritta e realizzazione di un **progetto di laboratorio** (project work), che deve essere svolto in gruppo (fino a tre studenti).

- La prova scritta ha l'obiettivo di valutare il livello di comprensione degli aspetti fondamentali dell'insegnamento e consiste in una serie di domande a risposta aperta.
- L'obiettivo del progetto di gruppo, tramite l'utilizzo di software open source, è lo sviluppo di soluzioni tecnologiche a problemi affrontati durante le lezioni. In particolare, si considerano ambiti applicativi reali che richiedono la definizione di sistemi i cui fondamenti sono stati presentati a lezione. Il progetto viene presentato di persona dagli studenti, per verificare le competenze effettivamente acquisite, sia di natura tecnica sia critica e di giudizio, e al contempo per sviluppare le capacità comunicative.

La **valutazione** della prova scritta avverrà in trentesimi. A fronte del conseguimento della sufficienza all'esame scritto (almeno 18/30), verranno aggiunti da 0 a 4 punti in base alla valutazione del progetto. Il progetto verrà valutato rispetto alla completezza rispetto alle specifiche, alla correttezza rispetto alle soluzioni algoritmiche impiegate, alla chiarezza di esposizione e alla qualità del materiale inviato.

Non sono previste prove in itinere.

Orario di ricevimento

Previo appuntamento con i docenti.

Sustainable Development Goals

SALUTE E BENESSERE | ISTRUZIONE DI QUALITÁ | PARITÁ DI GENERE | ENERGIA PULITA E ACCESSIBILE | IMPRESE, INNOVAZIONE E INFRASTRUTTURE | PACE, GIUSTIZIA E ISTITUZIONI SOLIDE