

## SCHEDA DELL'INSEGNAMENTO (SI)

TEXT MINING

SSD: ING-INF/05

DENOMINAZIONE DEL CORSO DI STUDIO: LAUREA MAGISTRALE IN INGEGNERIA INFORMATICA

ANNO ACCADEMICO 2023-2024

### INFORMAZIONI GENERALI - DOCENTE

DOCENTE: FLORA AMATO

TELEFONO: 081-7683851

EMAIL: FLORA.AMATO@UNINA.IT

### INFORMAZIONI GENERALI - ATTIVITÀ

INSEGNAMENTO INTEGRATO (EVENTUALE):

MODULO (EVENTUALE):

CANALE (EVENTUALE):

ANNO DI CORSO (I, II, III): I/II

SEMESTRE (I, II): I

CFU: 6

## **INSEGNAMENTI PROPEDEUTICI (se previsti dall'Ordinamento del CdS)**

**Nessuno**

## **EVENTUALI PREREQUISITI**

**Nessuno**

## **OBIETTIVI FORMATIVI**

Il corso ha l'obiettivo di fornire i principi metodologici e tecnologici relativi all'elaborazione di testi in linguaggio naturale. Questo prevede l'apprendimento dei principali algoritmi per l'analisi e l'interpretazione del linguaggio naturale, dei modelli utilizzati per la descrizione di argomenti e concetti, tecniche di estrazione delle informazioni e apprendimento automatico sui testi.

Gli studenti acquisiranno i principali modelli e le tecniche avanzate per l'annotazione automatica di risorse testuali, la formalizzazione della conoscenza contenuta nei testi, la strutturazione e l'estrazione di informazioni.

Relativamente all'apprendimento automatico, saranno approfondite ed utilizzate tecniche avanzate di Machine Learning e Deep Learning, applicate in diversi domini di interesse quali E-health, Giustizia, Amministrazione e Industria 4.0.

## **RISULTATI DI APPRENDIMENTO ATTESI (DESCRITTORI DI DUBLINO)**

### **Conoscenza e capacità di comprensione**

Il percorso formativo intende fornire agli studenti le conoscenze, i metodi e le tecniche necessarie per analizzare le problematiche sul trattamento automatico del testo, l'estrazione delle informazioni da esso, l'individuazione degli argomenti del discorso (topic).

Questi strumenti consentiranno agli studenti di capire le problematiche, padroneggiare la teoria, delineare una metodologia di strutturazione e analisi dei testi, di trattamento semantico e di rappresentazione della conoscenza alla base della costruzione di sistemi, e in particolare di interfacce intelligenti.

### **Capacità di applicare conoscenza e comprensione**

Il percorso formativo è orientato a trasmettere le capacità e gli strumenti metodologici e operativi necessari ad applicare le conoscenze di trattamento automatico del linguaggio naturale, nonché a favorire la capacità di utilizzare gli strumenti metodologici acquisiti per la progettazione e la realizzazione di sistemi e interfacce intelligenti. Le tecniche e i modelli proposti saranno applicati a domini specialistici quali E-health, Giustizia, Amministrazione e Industria 4.0.

## **PROGRAMMA-SYLLABUS**

- Introduzione alle tecniche di Elaborazione del Linguaggio Naturale (NLP). Intelligenza artificiale e NLP.
- Modelli di Rappresentazione delle Informazioni. Ontologie, vocabolari e dizionari personalizzati.
- Trasformazione di testo non strutturato in strutture dati idonee all'elaborazione automatica.
- Estrazione di caratteristiche testuali. Riconoscimento di modelli per scoprire e identificare categorie di informazioni.
- Apprendimento automatico ed elaborazione del linguaggio naturale.
- Algoritmi di apprendimento automatico di tipo Machine Learning per il testo.
- Pattern recognition per scoprire e identificare categorie di informazioni, confronto con approcci basati su dizionari.
- Identificazione, etichettatura e ricerca in sezioni specifiche del documento. Elaborazione linguistica per riconoscere unità significative all'interno del testo come frasi, gruppi di nomi, verbi e relazioni sussistenti.
- Strumenti semantici per identificare concetti all'interno del testo, normalizzazione dei concetti su ontologie standard. Individuazione di termini di dominio. Ontologie di dominio (come MedDRA e MeSH), dizionari personalizzati. identificazione, etichettatura e ricerca in sezioni specifiche del documento.
- Algoritmi di apprendimento automatico di tipo Deep Learning per il testo
- Reti neurali dense e ricorrenti, LSTMs, GRUs, in TensorFlow per eseguire l'analisi del testo, il riconoscimento delle entità e l'identificazione delle relazioni.
- Uso di encoder-decoder, causale, e auto-attenzione per eseguire l'elaborazione automatica avanzata di frasi complete, implementazione di interfacce intelligenti e chatbot.
- T5, BERT, transformer, reformer: uso dei modelli T5 e BERT per implementare un sistema automatico di risposta alle domande; implementazione di un chatbot usando un modello Reformer.

## MATERIALE DIDATTICO

### Libri di testo consigliati:

- Materiale prodotto e fornito dalla Docente
- D. Jurafsky, J.H. Martin, Speech and Language Processing. 3rd edition, Prentice-Hall, pp. 653, 2018 (con aggiornamenti di dicembre 2021 reperibili su [https://web.stanford.edu/~jurafsky/slp3/ed3book\\_dec292021.pdf](https://web.stanford.edu/~jurafsky/slp3/ed3book_dec292021.pdf)).
- S. Bird, E. Klein, E. Loper. Natural Language Processing with Python. O'Reilly, pp.903.

### Altro materiale didattico:

- Aggarwal, Charu C. Machine learning for text. Springer, 2018.
- Dipanjan Sarkar. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. Apress, 2019
- Clinical text mining: Secondary use of electronic patient records. Springer

## MODALITÀ DI SVOLGIMENTO DELL'INSEGNAMENTO

L'insegnamento si svolgerà con lezioni frontali (55% delle ore totali) ed esercitazioni di laboratorio (45% delle ore totali). Nelle esercitazioni di laboratorio sarà utilizzato il linguaggio Python, le librerie di trattamento testi NLTK e spaCy, i framework di deep learning TensorFlow e Keras. In modalità blended, tutte le lezioni saranno registrate e rese disponibili.

## VERIFICA DI APPRENDIMENTO E CRITERI DI VALUTAZIONE

### a) Modalità di esame:

L'esame si articola in prova	
scritta e orale	
solo scritta	
solo orale	X
discussione di elaborato progettuale	X
altro	

L'elaborato Progettuale sarà proposto al centro del corso e sarà sviluppato anche durante le lezioni di laboratorio con il supporto del docente.

### b) Modalità di valutazione:

La prova di esame avrà lo scopo di accertare il raggiungimento degli obiettivi formativi previsti per l'insegnamento, è articolata in una prova di laboratorio ed una prova orale incentrata sulla discussione del progetto.