UNIVERSITÀ DEGLI STUDI DI PALERMO

DIPARTIMENTO DI INGEGNERIA

ANNO ACCADEMICO PIANO DI STUDI	2022-2023
ANNO ACCADEMICO DI EROGAZIONE	2023-2024
CORSO DI LAUREA O DI LAUREA MAGISTRALE	Laurea Magistrale in Ingegneria Informatica – Intelligenza Artificiale LM 32
INSEGNAMENTO	Elaborazione del Linguaggio Naturale
TIPO DI ATTIVITÀ	Caratterizzante
AMBITO DISCIPLINARE	Ingegneria Informatica
CODICE INSEGNAMENTO	16554
ARTICOLAZIONE IN MODULI	NO
NUMERO MODULI	1
SETTORI SCIENTIFICO DISCIPLINARI	ING-INF/05
DOCENTE RESPONSABILE (MODULO 1)	Roberto Pirrone Professore Associato Università degli Studi di Palermo roberto.pirrone@unipa.it
CFU	6
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	96
NUMERO DI ORE RISERVATE ALLE ATTIVITÀ DIDATTICHE ASSISTITE	54
PROPEDEUTICITÀ	Nessuna
PREREQUISITI	Conoscenze di Machine Learning acquisite nell'insegnamento di Big data C.I. erogato al I anno di corso.
ANNO DI CORSO	secondo
SEDE DI SVOLGIMENTO DELLE LEZIONI	Aula da definire
ORGANIZZAZIONE DELLA DIDATTICA	Lezioni frontali; Esercitazioni teoriche; Esercitazioni di gruppo per lo sviluppo di una pipeline di elaborazione del linguaggio naturale per rispondere ad un caso di studio proposto dal docente.
MODALITÀ DI FREQUENZA	Facoltativa
METODI DI VALUTAZIONE	L'esame finale consta di un colloquio orale in cui, oltre a valutare

	la conoscenza dei temi teorici affrontati durante il corso, si discuteranno i risultati e le soluzioni utilizzate per implementare il caso di studio. Il colloquio orale tende ad approfondire le conoscenze e le competenze pratiche riguardo ai temi presenti nel programma. Gli aspetti dell'esposizione che saranno valutati sono: • Grado di comprensione mostrato in relazione al programma teorico svolto • Proprietà del linguaggio utilizzato • Capacità di giudizio autonomo sull'utilizzo dei diversi algoritmi affrontati in relazione alle differenti esigenze legate ai task di elaborazione del linguaggio naturale • Capacità di problem solving La presentazione del caso di studio verrà valutata secondo i seguenti aspetti del codice prodotto: • Completezza • Originalità • Grado di performance ottenuto in termini delle metriche di valutazione assegnate per il caso di studio • Organizzazione del codice prodotto • Capacità di integrazione di codice già noto dalle esercitazioni teoriche. L'articolazione del voto di esame sarà strutturata per fasce di valutazione: - 18/30 – 20/30: lo studente ha una conoscenza appena sufficiente dei contenuti teorici dell'insegnamento ed è in grado di sviluppare solo alcune parti della pipeline che affronta il caso di studio. - 21/30 – 23/30: lo studente ha una discreta conoscenza dei contenuti teorici dell'insegnamento; egli riesce a sviluppare sommariamente l'intera pipeline senza curarsi delle problematiche legate sia al pre-processing dei dati sia all'addestramento del modello. - 24/30 – 26/30: lo studente ha buona conoscenza dei contenuti teorici dell'insegnamento e sviluppa l'intera pipeline dimostrando di analizzare almeno in parte i problemi connessi sia al pre-processing dei dati sia all'addestramento del modello. - 27/30 – 30/30: lo studente ha piena conoscenza dell'insegnamento e sviluppa completamente e correttamente la pipeline richiesta dal caso di studio analizzando tutti gli aspetti connessi sia al pre-processing dei dati sia all'addestramento del modello. - 30 e lode: lo studente conosce ottimamente gli argomenti teorici del corso e ha ottime capacità di sviluppo della pipeline proposta come caso di studio insieme a tutti gli aspetti a questa connessi; egli inoltre mostra originalità e capacità di approfondimento autonomo dei temi trattati: le sue soluzioni di analisi del linguaggio naturale sono altresì innovative.
TIPO DI VALUTAZIONE	Voto in trentesimi
PERIODO DELLE LEZIONI	Secondo semestre

CALENDARIO DELLE ATTIVITÀ DIDATTICHE	-
ORARIO DI RICEVIMENTO DEGLI STUDENTI	Il mercoledì dalle 11.30 alle 13.00, salvo impegni istituzionali, presso il Dipartimento di Ingegneria (DI) Ed. 6, III piano, stanza 3025 e/o sul team dedicato con codice: 4rylimr .

RISULTATI DI APPRENDIMENTO ATTESI

Conoscenza e capacità di comprensione
Lo studente, al termine del corso, avrà acquisito conoscenze e metodologie per affrontare le problematiche legate ai modelli predittivi di elaborazione del linguaggio naturale. Egli conoscerà i principali ambiti applicativi della disciplina e le diverse tipologie di algoritmi utilizzati in ciascuno di essi.
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 discussione orale.

Capacità di applicare conoscenza e comprensione
Lo studente avrà acquisito conoscenze e metodologie per analizzare e risolvere problemi tipici legati allo sviluppo di una pipeline di elaborazione dei testi per affrontare un problema di elaborazione del linguaggio naturale.
Egli avrà conoscenza approfondita degli ambienti software basati sul linguaggio Python più diffusi per l'elaborazione del linguaggio naturale quali NLTK e PyTorch.
Per il raggiungimento di quest'obiettivo il corso comprende: esercitazioni teoriche e di gruppo su un caso di studio proposto dal docente per lo sviluppo di modelli predittivi per l'elaborazione del linguaggio naturale.
Per la verifica di quest'obiettivo l'esame comprende la discussione delle soluzioni adottate dai diversi gruppi nel risolvere il caso di studio.

Autonomia di giudizio
Lo studente sarà in grado di svolgere un'analisi comparativa delle caratteristiche dei differenti algoritmi di elaborazione del linguaggio naturale. Egli sarà 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 potrà condurre lo sviluppo di nuove problematiche applicative nell'ambito dell'elaborazione del linguaggio naturale.
Per il raggiungimento di quest'obiettivo il corso comprende le esercitazioni guidate sullo sviluppo del caso di studio proposto dal docente.
Per la verifica di quest'obiettivo l'esame comprende la discussione sulle caratteristiche dell'implementazione proposta dal gruppo di lavoro.

Abilità comunicative
Lo studente sarà in grado di comunicare con competenza e proprietà di linguaggio problematiche complesse di elaborazione del linguaggio naturale in contesti specializzati.
Per il raggiungimento di quest'obiettivo il corso comprende le esercitazioni guidate sullo sviluppo del caso di studio proposto dal docente.
Per la verifica di quest'obiettivo l'esame comprende la discussione sulle caratteristiche dell'implementazione proposta dal gruppo di lavoro.

Capacità d'apprendimento
Lo studente sarà in grado di affrontare in autonomia qualsiasi problematica concernente lo sviluppo di

modelli predittivi di elaborazione del linguaggio naturale. Sarà in grado di approfondire autonomamente le tematiche complesse legate allo sviluppo di algoritmi innovativi nei diversi contesti applicativi del NLP.
Per il raggiungimento di quest'obiettivo il corso comprende le esercitazioni guidate sullo sviluppo del caso di studio proposto dal docente.
Per la verifica di quest'obiettivo l'esame comprende la discussione sulle caratteristiche dell'implementazione proposta dal gruppo di lavoro ed il colloquio orale in cui verranno verificate le capacità di problem solving dell'allievo.

OBIETTIVI FORMATIVI DELL'INSEGNAMENTO

Il corso di "Elaborazione del Linguaggio Naturale" fornisce agli studenti una conoscenza approfondita degli algoritmi, dei linguaggi e dei tipici ambienti software per lo sviluppo di pipeline di addestramento di modelli per il Natural Language Processing.

Il corso consente di acquisire 6 CFU e consta di una serie di lezioni ed esercitazioni teoriche nonché la costituzione di gruppi di lavoro per l'analisi di un caso di studio proposto dal docente attraverso lo sviluppo di un modello di Machine Learning/Deep Learning. Il risultato dell'attività inizialmente guidata in aula e poi autonoma dei gruppi di lavoro viene poi discusso durante il colloquio orale.

Il ciclo di lezioni teoriche presenta dapprima una introduzione alle tecniche di pre-processing dei dati testuali per la creazione di data set. Successivamente vengono presentate brevemente le tecniche probabilistiche per il Natural Language Processing: i modelli del linguaggio basati su N-grammi e le tecniche di Part Of Speech Tagging e Named Entity Recognition. Si passa poi ad illustrare le tecniche basate su reti neurali: gli embedding non contestuali come word2vec, i modelli del linguaggio basati su reti ricorrenti e transformer, il meccanismo di attenzione con la traduzione automatica e l'utilizzo dei grandi modelli del linguaggio pre-addestrati. Infine, viene presentata una rassegna dei principali campi di applicazione: il parsing e il concetto di grammatica, l'estrazione di relazioni, l'uso di lessici come WordNet, la sentiment analysis, la coreference resolution e il question answering.

Le esercitazioni teoriche coprono la configurazione degli ambienti di sviluppo con cui si opererà durante il corso, quali la libreria Python NLTK e PyTorch per l'implementazione di modelli neurali, nonché l'illustrazione dei temi affrontati nel corso teorico attraverso esempi svolti.

Infine, i gruppi di lavoro svolgeranno, ciascuno separatamente, il caso di studio proposto dal docente, dapprima in una serie di esercitazioni guidate e poi in autonomia per la preparazione dell'esame.

Elaborazione del Linguaggio Naturale	
ORE FRONTALI	**LEZIONI FRONTALI**
1	Introduzione al Corso
2	Processing del testo: tokenizzazione, segmentazione, normalizzazione, lemmatizzazione, stemming
2	Modelli del linguaggio con N-grammi
4	Part-of-Speech Tagging e Named Entity Recognition
2	Semantica vettoriale ed embedding non contestuali: word2vec, GloVE, fasttext
4	Modelli neurali del linguaggio: reti feed-forward, reti ricorrenti e meccanismo di attenzione
3	Traduzione automatica: il meccanismo di self-attention e i Transformer
3	Masked Language Models pre-addestrati: BERT, RoBERTa e derivati
3	Large Language Models: ChatGPT e LLAMA2, prompting, RAG e fine-tuning

ORE FRONTALI	**ESERCITAZIONI TEORICHE**
3	Introduzione a PyThorch
3	Text processing con NLTK e modelli a N-grammi
3	POS e NER con NLTK
3	Embedding e reti neurali per la text classification
3	Trasformer per la traduzione automatica con la piattaforma Hugging Face
3	Uso di BERT da Hugging Face per i task NLP già illustrati in precedenza
3	Uso di RoBERTa e modelli multilingua da Hugging Face per lo svolgimento di task in italiano
3	Uso di LLAMA2-7b da Hugging Face per lo svolgimento di task NLP

ORE FRONTALI	**ALTRO**
6	Assegnazione e svolgimento della challenge legata al caso di studio

TESTI	Daniel Jurafsky, James H. Martin “Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”, 3d Edition draft, 2023, disponibile on line all’indirizzo: https://web.stanford.edu/~jurafsky/slp3/ Steven Bird, Ewan Klein, and Edward Loper, “Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit” versione on line all’indirizzo: https://www.nltk.org/book/ Materiale didattico in forma elettronica disponibile sul repository GitHub predisposto dal docente: https://github.com/fredffsixty/Natural_Language_Processing