UNIVERSITÀ DI PISA

TEXT ANALYTICS

LUCIA PASSARO

Anno accademico	2022/23
CdS	DATA SCIENCE AND BUSINESS INFORMATICS
Codice	635AA
CFU	6

Moduli	Settore/i	Tipo	Ore	Docente/i
TEXT ANALYTICS	INF/01	LEZIONI	48	LUCIA PASSARO

Obiettivi di apprendimento

Conoscenze
Apprendimento di tecniche, algoritmi e modelli essenziali utilizzati nell'elaborazione del linguaggio naturale. Comprensione delle architetture delle tipiche applicazioni di analisi del testo e delle librerie per la loro realizzazione. Competenza nella progettazione, implementazione e valutazione di applicazioni che sfruttano l'analisi, l'interpretazione e la trasformazione dei testi.

Modalità di verifica delle conoscenze
Lo studente sarà valutato in base alla sua capacità di discutere i contenuti del corso utilizzando la terminologia appropriata e di applicare le tecniche di elaborazione del linguaggio naturale.

Capacità
Lo studente sarà in grado di progettare, implementare e valutare applicazioni basate sull'analisi, l'interpretazione e la trasformazione dei testi.

Modalità di verifica delle capacità
Agli studenti frequentanti verrà chiesto di partecipare a un progetto di gruppo volto a valutare le competenze nella progettazione e nell'implementazione di un compito di analisi del testo concordato con il docente.
Agli studenti non frequentanti verrà chiesto di risolvere esercizi durante un esame scritto e una discussione orale.

Comportamenti
Gli studenti saranno in grado di analizzare un problema di elaborazione del testo, selezionare i metodi corretti per risolverlo e implementare una soluzione funzionante. Saranno consapevoli di diversi problemi legati all'elaborazione del linguaggio naturale, tra cui l'affidabilità dei risultati, quando le applicazioni coinvolgono dati (soggettivi) annotati dall'uomo.

Modalità di verifica dei comportamenti
Il comportamento degli studenti sarà valutato durante lo sviluppo del progetto e/o all'esame scritto/orale.

Prerequisiti (conoscenze iniziali)
Prerequisiti utili:

- Coding (python)
- Probability theory
- Information theory

Indicazioni metodologiche
Modalità: lezioni frontali in lingua inglese
Attività:

- partecipazione alle lezioni
- partecipazione ai seminari tenuti da aziende e/o esperti della materia
- partecipazione alle discussioni

- studio individuale
- esercizi (con tool gratuiti)
- progetto di gruppo

Frequenza: fortemente consigliata
Metodi di insegnamento:

- Lezioni
- Seminari tenuti da aziende e/o esperti della materia

Saranno presentati casi di studio settoriali, possibilmente durante i seminari, con la partecipazione attiva degli studenti.

Programma (contenuti dell'insegnamento)

1. Background: Elaborazione del linguaggio naturale, recupero delle informazioni e apprendimento automatico.
2. Background matematico: Probabilità, statistica e algebra
3. Elementi linguistici essenziali: parole, lemmi, morfologia, parte del discorso (PoS), sintassi
4. Elaborazione di base del testo: espressione regolare, tokenizzazione
5. Raccolta dati: scraping
6. Modellazione: collocazioni, modelli linguistici
7. Introduzione al Machine Learning: teoria e suggerimenti pratici
8. Librerie e strumenti: NLTK, Spacy, Keras, pytorch
9. Classificazione/Clustering
10. Analisi del sentimento/estrazione di opinioni
11. Estrazione di informazioni/estrazione di relazioni/collegamento di entità
12. Ttransfer learning
13. Quantification

Bibliografia e materiale didattico

E' raccomandata la lettura di capitoli selezionati tratti da:

1. D. Jurafsky, J.H. Martin, Speech and Language Processing. 3nd edition, Prentice-Hall, 2018.
2. S. Bird, E. Klein, E. Loper. Natural Language Processing with Python.

Bibliografia aggiuntiva sarà indicata sulla pagina web del corso.

Indicazioni per non frequentanti

Gli studenti non frequentanti non possono svolgere il progetto. L'esame consisterà in una prova scritta con domande aperte ed esercizi e in una discussione orale sugli argomenti del corso.

Modalità d'esame

L'esame consiste in una parte scritta e in una parte orale. La parte scritta dura 2 ore e comprende domande aperte ed esercizi. A ogni esercizio viene assegnato un punteggio. Gli studenti sono ammessi alla parte orale se il loro punteggio totale è di almeno 18/30. La parte orale consiste in domande aperte sugli argomenti del corso e sull'uso di strumenti per l'analisi del testo.
Gli studenti frequentanti possono sostituire la parte scritta con un progetto da svolgere in gruppo durante il corso. Il risultato del progetto sarà del codice e una relazione sull'attività svolta (lunghezza tipica: 4-10 pagine). L'esame orale consisterà nella presentazione e nella discussione del progetto. Durante l'esame orale sarà valutato il contributo individuale dello studente al progetto di gruppo.

Pagina web del corso

http://didawiki.cli.di.unipi.it/doku.php/mds/txa/start

Ultimo aggiornamento 14/09/2022 12:46