UNIVERSITÀ DI PISA

LINGUISTICA COMPUTAZIONALE II

GIULIA VENTURI

Anno accademico	2023/24
CdS	INFORMATICA UMANISTICA
Codice	513LL
CFU	6

Moduli	Settore/i	Tipo	Ore	Docente/i
LINGUISTICA COMPUTAZIONALE II	L-LIN/01	LEZIONI	36	FELICE DELL'ORLETTA SIMONETTA MONTEMAGNI GIULIA VENTURI

Obiettivi di apprendimento

Conoscenze
Il corso si propone di introdurre lo studente a settori chiave della Linguistica Computazionale caratterizzati da un forte impatto applicativo. In particolare, si articola attorno a due macro-temi: 1) metodi e strumenti di annotazione linguistica multi-livello del testo per l'estrazione di conoscenza linguistica da corpora testuali e 2) modelli di classificazione del testo basati su caratteristiche linguistiche esplicite e rappresentazioni distribuite delle parole. Entrambe le tematiche sono affrontate da una duplice prospettiva, teorica e applicativa.

Modalità di verifica delle conoscenze
Svolgimento a scelta tra due progetti; relazione scritta sui risultati del progetto scelto, da presentare in fase di iscrizione all'esame orale; esame orale in cui verranno discussi i risultati del progetto e verificata la conoscenza dei temi trattati durante il corso.

Capacità
Al termine del corso lo studente saprà a) utilizzare in modo critico e consapevole strumenti di annotazione linguistica automatica e di estrazione di conoscenza linguistica, b) sviluppare modelli di classificazione automatica del testo, c) identificare le problematiche legate al trattamento di varietà non-standard della lingua e ipotizzare possibili soluzioni.

Prerequisiti (conoscenze iniziali)
Nozioni di base di linguistica computazionale, di linguistica generale e di linguistica italiana. E' fortemente consigliato aver frequentato e sostenuto l'esame di Linguistica Computazionale I.

Indicazioni metodologiche
Durante il corso si alterneranno lezioni frontali, con l'ausilio di slides powerpoint che vengono messe a disposizione degli studenti, ed esercitazioni di laboratorio, sia individuali sia di gruppo (svolte con PC delle aule informatiche e/o PC personali), in cui gli studenti sono invitati a confrontarsi con l'applicazione di strumenti software di annotazione linguistica del testo e di estrazione di conoscenza disponibili come demo online, nonché con lo sviluppo di modelli di classificazione automatica del testo. Gli studenti saranno inoltre chiamati ad analizzare criticamente i risultati di annotazione e classificazione ottenuti in relazione a diverse varietà d'uso della lingua.

Programma (contenuti dell'insegnamento)
I contenuti del programma potranno subire variazioni e/o integrazioni, che verranno comunicate durante la prima lezione del corso.
I contenuti del corso sono suddivisi in due macro-temi, per ciascuno dei quali segue una lista dei principali argomenti trattati:

- **Annotazione linguistica**
  - annotazione linguistica come processo incrementale; strumenti software per l'annotazione linguistica del testo; schemi di annotazione per l'annotazione morfo-sintattica e sintattica, con particolare attenzione allo schema delle "Universal Dependencies"; costruzione di corpora annotati e valutazione dell'annotazione; adattamento al dominio o altre varietà d'uso della lingua di strumenti di annotazione;
- **Estrazione di conoscenza linguistica**
  - ricostruzione del profilo linguistico di collezioni di testi; monitoraggio linguistico di diverse tipologie testuali e/o varietà d'uso della lingua; uso dei risultati del monitoraggio linguistico all'interno di diversi scenari applicativi, ad esempio

per la classificazione di generi testuali o per l'identificazione della lingua materna di produzioni L2; analisi della leggibilità del testo.

- **Modelli di classificazione del testo**:
  - sviluppo e valutazione di algoritmi per la classificazione di documenti utilizzando le librerie scipy e scikit-learn
- **Impiego di caratteristiche linguistiche esplicite e rappresentazioni distribuite delle parole per lo sviluppo di modelli di classificazione del testo**:
  - rappresentazione del testo attraverso caratteristiche linguistiche esplicite: dal lessico alle informazioni sintattiche
  - rappresentazione del testo attraverso rappresentazioni distribuite non contestuali (word embedding), contestuali (contestual word embedding) e modelli del linguaggio.

Bibliografia e materiale didattico

L'elenco dei testi d'esame è disponibile alla pagina Moodle del corso. Gli studenti non frequentanti sono pregati di contattare i docenti per concordare il programma d'esame.

Indicazioni per non frequentanti

Dato il carattere estremamente applicativo del corso, la frequenza è fortemente richiesta. Qualora lo studente fosse impossibilitato a frequentare, si prega di contattare i docenti per concordare il programma d'esame.

Modalità d'esame

Le modalità d'esame verranno comunicate durante la prima lezione del corso.

Ultimo aggiornamento 04/08/2023 17:25