UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Natural Language Processing

2627-2-FDS02Q011

Obiettivi

Lo scopo del corso è fornire un'introduzione ai concetti fondamentali relativi all'elaborazione del linguaggio naturale (NLP) nonché una panoramica dei principali strumenti utilizzati nel settore. Inoltre, verranno presentate alcune applicazioni di NLP, ad esempio traduzione automatica e rilevamento di discorsi di odio.

Gli obiettivi specifici rispetto ai Descrittori di Dublino (DdD) sono i seguenti:
1. Conoscenza e capacità di comprensione

- Acquisire familiarità con i concetti fondamentali del Natural Language Processing (NLP).
- Comprendere le principali tecniche statistiche e neurali per l'elaborazione del linguaggio naturale.
- Conoscere gli strumenti e modelli più recenti nel settore del NLP, inclusi Word2Vec, FastText, GloVe, BERT, GPT.

2. Conoscenza e capacità di comprensione applicate

- Applicare algoritmi e modelli NLP a problemi reali quali machine translation e hate speech detection.
- Utilizzare strumenti open-source per implementare sistemi di NLP.
- Progettare e sviluppare un'applicazione NLP partendo da un dominio specifico

3. Autonomia di giudizio

- Essere in grado di scegliere i metodi e modelli più appropriati per risolvere compiti specifici relativamente all'analisi del linguaggio naturale.
- Analizzare l'efficacia di modelli linguistici

4. Abilità comunicative

- Presentare in modo chiaro e argomentato un progetto di NLP, anche attraverso esposizione orale.
- Comunicare i risultati dell’analisi a un pubblico tecnico e non.

- Collaborare in gruppo per la realizzazione di un progetto complesso, condividendo conoscenze e strategie.

5. Capacità di apprendere

- Sviluppare capacità autonome di studio attraverso il completamento di assignment facoltativi derivanti da attività di laboratorio focalizzate all'uso di modelli e strumenti di NLP.

Contenuti sintetici

Il contenuto del corso include i principi fondamentali dell'elaborazione del linguaggio naturale (NLP) e offre una panoramica degli strumenti chiave utilizzati in questo campo. Il corso coprirà una vasta gamma di argomenti, che vanno dalle tecniche statistiche ai recenti progressi negli approcci neurali. Inoltre, il corso comprende dimostrazioni pratiche di diverse applicazioni di NLP, tra cui traduzione automatica e rilevamento dell'incitamento all'odio.

Programma esteso

Introduzione al corso
- Approcci razionalisti ed empiristi al linguaggio
- L'ambiguità del linguaggio: perché il NLP è difficile

Elementi linguistici essenziali
- Risorse lessicali
- Legge di Zipf
- Collocazioni
- Sintassi

Rappresentazione frequentista del testo (TF, TF-IDF, ecc.) e word embedding
- Word2Vec
- FastText
- Glove

Tecniche di visualizzazione degli embeddings:
- PCA
- t-SNE
- UMAP

Sequence-to-sequence (RNN, LSTM)

Transformer e modelli linguistici di grandi dimensioni
- Meccanismi di attenzione: self- e multi-head attention

Modelli linguistici contestualizzati:
- ELMO
- BERT
- GPT
- LLAMA

Prompting e fine-tuning

Metriche di valutazione dei modelli del linguaggio

Explainability dei modelli linguistici

Prerequisiti

Basic knowledge of statistics and programming languages.

Modalità didattica

Il corso sarà tenuto in lingua inglese e si articolerà sia in lezioni frontali che introdurranno gli argomenti principali, sia in sessioni tutoriali in cui verranno spiegati gli strumenti open source.
Possono far parte del corso seminari tenuti da esperti a livello nazionale e internazionale.

24 lezioni da 2 ore svolte in modalità erogativa in presenza.

Materiale didattico

Daniel Jurafsky and James Martin, "Speech and Language Processing, 2nd Edition", Prentice Hall, 2008.

Emily M. Bender, "Linguistic Fundamentals for Natural Language Processing", Synthesis lectures on human language technologies, Morgan&Claypool Publishers, 2013.

Yoav Goldberg, "Neural Network Methods for Natural Language Processing", Synthesis lectures on human language technologies, Morgan&Claypool Publishers, 2017.

Mohammad Taher Pilehvar and Jose Camacho-collados, "Embeddings in Natural Language Processing", Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2021.

Periodo di erogazione dell'insegnamento

Second Semester

Modalità di verifica del profitto e valutazione

Progetto
• Il progetto consiste nello sviluppo di uno strumento per l'elaborazione del linguaggio naturale basato su metodi e modelli presentati durante il corso.
• E' necessario identificare un dominio di interesse e un set di dati per il quale si intende affrontare compiti specifici che coinvolgano metodi, modelli e strumenti di NLP.
• Il progetto deve essere presentato oralmente
• Il progetto viene valutato nell'intervallo [0-24] in accordo ai seguenti criteri:

1. Definizione del problema e degli obiettivi: max 2 punti
2. Giustificazione delle scelte metodologiche: max 3 punti
3. Implementazione e correttezza del codice: max 4 punti
4. Analisi dei risultati sperimentali: max 6 punti
5. Completezza e chiarezza della relazione finale: max 6 punti
6. Conclusioni e considerazioni finali: max 3 punti

Esame orale
• La prova orale può avere un esito compreso tra [-8; +8]
• Si compone di 4 domande su argomenti affrontati durante il corso: -2 punti per una risposta errata o per nessuna risposta, +2 punti per una risposta corretta.

Non sono previste prove in itinere.

Orario di ricevimento

To be agreed with the teacher

Sustainable Development Goals