UNIVERSITÀ DI PAVIA

Anno Accademico 2021/2022

LINGUISTICA COMPUTAZIONALE

Anno immatricolazione	2021/2022
Anno offerta	2021/2022
Normativa	DM270
SSD	L-LIN/01 (GLOTTOLOGIA E LINGUISTICA)
Dipartimento	DIPARTIMENTO DI STUDI UMANISTICI
Corso di studio	LINGUISTICA TEORICA, APPLICATA E DELLE LINGUE MODERNE
Curriculum	PERCORSO COMUNE
Anno di corso	1°
Periodo didattico	Primo Semestre (27/09/2021 - 23/12/2021)
Crediti	6
Ore	36 ore di attività frontale
Lingua insegnamento	Italiano
Tipo esame	SCRITTO E ORALE CONGIUNTI
Docente	JEZEK ELISABETTA (titolare) - 6 CFU
Prerequisiti	Nozioni di base di linguistica, che saranno riprese all'inizio del corso.
Obiettivi formativi	L'analisi automatica dei testi è oggi essenziale per scopi di ricerca nelle scienze umane e sociali e per applicazioni di varia natura, dalla traduzione automatica, alla estrazione di opinioni, alla costruzione di agenti conversazionali. Il corso introduce i concetti, le metodologie e gli strumenti fondamentali della linguistica computazionale e del trattamento automatico del linguaggio, fornendo agli studenti competenze per analizzare automaticamente o semiautomaticamente dati testuali di varia natura (letterari, storici, scientifici, socio-politici, giornalistici). Sono inoltre fornite le basi metodologiche dell'annotatione linguistica dei testi per l'apprendimento automatico supervisionato.
Programma e contenuti	Il corso costituisce una introduzione ai fondamenti della linguistica computazionale e del trattamento automatico dei testi.

Coprirà i seguenti argomenti:

- Elementi di linguistica per l'analisi computazionale dei testi
- Basi di Statistica
- Il trattamento automatico del testo
- Metodi di apprendimento automatico
- La annotazione di dati linguistici per l'apprendimento automatico
- I principali task nel trattamento automatico dei testi (con approfondimento sul riconoscimento automatico dei nomi propri - Named Entity Recognition, sull'identificazione di informazioni temporali e tipi di eventi - Temporal Information Extraction e Event Detection - e sull'estrazione di opinioni - Opinion Mining e Sentiment Analysis).

Due lezioni di carattere interdisciplinare verteranno sul tema "Machine Learning for the Social Sciences and the Humanities" e si terranno in inglese.

Il corso includerà una parte laboratoriale con esercitazioni relative all'analisi automatica di testi. Gli studenti acquisiranno competenze relative all'uso dell'interfaccia a linea di comando per la manipolazione dei testi e di alcuni strumenti automatici per l'estrazione di informazioni linguistiche. In particolare verranno introdotti i seguenti strumenti: le pipeline UDpipe e Tint, uno script per la Sentiment Analysis basata sul lessico e le funzioni di base del Natural Language Processing ToolKit (NLTK). Quest'ultimo richiede nozioni di base di programmazione in Python, che gli studenti dovranno rapidamente acquisire in classe durante la prima settimana del corso.

Metodi didattici

Lezioni frontali interattive.
Slides.
Laboratorio con esercitazioni.

Testi di riferimento

Letture:

Jezek, Elisabetta 2016. The Lexicon: An Introduction, Oxford, Oxford University Press. Capitolo 1 "Basic Notions".

Jurafsky, Dan, and James H. Martin. 2018. Speech and language processing. Ed. 3. URL: https://web.stanford.edu/~jurafsky/slp3/17.pdf - Capitolo 17, Sezione 17.1 "Named Entity Recognition".

Jurafsky, Dan, and James H. Martin. 2018. Speech and language processing. Ed. 3. URL: https://web.stanford.edu/~jurafsky/slp3/17.pdf - Capitolo 6, "Vector Semantics".

Liu, B. 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167. Capitolo 1 "Sentiment Analysis: A Fascinating Problem" e Capitolo 2 "The Problem of Sentiment Analysis". Disponibile tramite Linkup, Università di Pavia, https://www.morganclaypool.com/action/ssostart?redirectUri=%2F.

Lu, X., 2014. Computational methods for corpus annotation and analysis. Dordrecht, Springer. Capitolo 2 "Text Processing with the

Command Line Interface".

Straka, Milan, Jan Hajic, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of LREC 2016. URL: http://ufal.mff.cuni.cz/~straka/papers/2016-lrec_udpipe.pdf

Pustejovsky, James, José M. Castano, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. New directions in question answering 3: 28-34. URL: https://www.aaai.org/Papers/Symposia/Spring/2003/SS-03-07/SS03-07-005.pdf

Pustejovsky J. and A. Stubbs. 2012. Natural Language Annotation for Machine Learning, 0'Reylly Media, Capitolo 7 Training: Machine Learning.

Natural Language ToolKit. URL: https://towardsdatascience.com/introduction-to-natural-language-processing-for-text-df845750fb63

Opzionale:

Hürriyeto?lu, A., Zavarella, V., Tanev, H., Yörük, E., Safaya, A. and Mutlu, O., 2020. Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. Proceedings of AESPEN 2020, Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020, p. 1-6. https://www.aclweb.org/anthology/2020.aespen-1.1.pdf

Ulteriori letture saranno indicate durante le lezioni e indicate nella piattaforma KIRO.

Modalità verifica apprendimento

Prova orale di verifica dell’apprendimento dei contenuti del corso. Indagine empirica di un fenomeno linguistico (sintattico, semantico, lessicale, discorsivo) o di un fenomeno storico, culturale, sociale attraverso l'analisi linguistica, a scelta dello studente, concordato con la docente, utilizzando gli strumenti di analisi automatica dei testi illustrati nel corso.
Elaborato scritto di 5 cartelle (inclusa bibliografia, escluse le tabelle e le figure) riportante i risultati del task svolto, da inviare a jezek@unipv.it 7 gg prima della data dell'appello d’esame.

Altre informazioni

Tutto il materiale didattico è disponibile sul portale della didattica KIRO (accesso con credenziali di Ateneo).

Obiettivi Agenda 2030 per lo sviluppo sostenibile

Questo insegnamento concorre alla realizzazione dei 17 “Sustainable Development Goals” (obiettivi di sviluppo sostenibile) delle Nazioni Unite per il 2030, specificamente il n. 4: "Quality Education", che è "mirato a garantire un'istruzione di qualità inclusiva ed equa e a promuovere opportunità di apprendimento permanente per tutti".

https://unric.org/it/agenda-2030/
$lbl_legenda_sviluppo_sostenibile