UNIVERSITÀ DI PISA

TECNOLOGIE LINGUISTICHE PER L'ESTRAZIONE DI INFORMAZIONE

FEDERICO BOSCHETTI

Anno accademico	2017/18
CdS	INFORMATICA UMANISTICA
Codice	569LL
CFU	6

Moduli	Settore/i	Tipo	Ore	Docente/i
TECNOLOGIE LINGUISTICHE PER L'ESTRAZIONE DI INFORMAZIONE	L-LIN/01	LEZIONI	42	FEDERICO BOSCHETTI

Obiettivi di apprendimento

Conoscenze

- Apprendimento delle nozioni teoriche relative al trattamento delle immagini digitali di pagine di testo, al riconscimento ottico dei caratteri (OCR) e alla correzione semiautomatica del testo acquisito.
- Conoscenza delle diverse teorie riguardanti l'edizione scientifica digitale.
- Apprendimento delle principali tecniche di analisi linguistica e stilistica applicata a testi di interesse storico-letterario.

Modalità di verifica delle conoscenze
Viene valutata l'acquisizione delle conoscenze tramite colloquio.

Capacità
Capacità di seguire il flusso di lavoro dall'acquisizione del testo tramite OCR, attraverso la creazione dell'edizione digitale, fino all'analisi testuale.

Modalità di verifica delle capacità
Lo studente concorda con il docente un progetto di digitalizzazione tramite OCR, marcatura semiautomatica del testo ed annotazione linguistica e/o stilistica. Viene valutata la relazione scritta che illustra il progetto.

Comportamenti
Stesura di un progetto di digitalizzazione tramite OCR, creazione dell'edizione digitale e relative analisi linguistiche e stilistiche.

Modalità di verifica dei comportamenti
Colloqui e revisioni del progetto.

Prerequisiti (conoscenze iniziali)
Conoscenze informatiche di base.

Indicazioni metodologiche

- Le lezioni frontali si svolgono prevalentemente con l'ausilio di slides;
- I materiali didattici sono messi progressivamente a disposizione sulla piattaforma moodle del Polo 4;
- Le esercitazioni pratiche si svolgono prevalentemente con l'uso dei portatili personali degli studenti.

Programma (contenuti dell'insegnamento)
Introduzione

- Introduzione generale

Acquisizione

- Scanner e repertori di immagini
- Trattamento delle immagini
- Optical Character Recognition (OCR)
- Algoritmi di allineamento
- Tecniche linguistiche per migliorare l'accuratezza
- Applicazioni per la correzione collaborativa dell'OCR

Edizioni e Annotazioni

- Che cos'è un'edizione digitale
- Canonical Texts Services (CTS)
- Edizione critica e Text Encoding Initiative (TEI)
- Rappresentazione della variantistica
- Piattaforme web per l'annotazione
- Annotazione tramite Domain Specific Languages
- Dalle folksonomies alle ontologie

Analisi

- Lemmatizzazione e analisi morfologica di testi antichi
- Treebanking: varianti e interpretazioni
- Semantica distribuzionale diacronica
- Named Entity Recognition in prospettiva diacronica
- Analisi metrica
- Elementi di Stilometria

Conclusione

- Discussione generale sui risultati raggiunti

Bibliografia e materiale didattico

- Driscoll, Matthew James, and Elena Pierazzo (eds.). 2016. *Digital scholarly editing: theories and practices*. Cambridge, UK: Open Book Publishers. http://dx.doi.org/10.11647/OBP.0095
- Piotrowski, Michael. 2012. *Natural Language Processing for Historical Texts*. San Rafael: Morgan & Claypool Publishers.
- Schreibman, Susan, Ray Siemens, and John Unsworth (eds.). 2016. *A new companion to digital humanities*. Chichester: Wiley Blackwell.

Ulteriori informazioni e materiali didattici saranno forniti durante il corso.

Indicazioni per non frequentanti

I non frequentanti sono invitati a concordare il programma con il docente.

Modalità d'esame

Relazione scritta sul progetto e interrogazione orale sulla parte teorica.

Stage e tirocini

È possibile svolgere il tirocinio presso l'Istituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa.

Ultimo aggiornamento 18/08/2017 17:51