

Corso di *Web Mining e Retrieval*

- Introduzione al Corso -

Corso di Laurea in Informatica
(a.a. 2012-2013)

Roberto Basili



Overview

- WM&R: Motivazioni e prospettive
- Modalità di erogazione del Corso
- Prerequisiti
- Forma e struttura delle prove d'esame

Obbiettivi del Corso

- Metodi per l'accesso ed il trattamento dell'informazione distribuita
- Componente Fondazionale:
 - Problemi di Ricerca e Classificazione dei testi
 - Modelli Avanzati di *Machine Learning*
 - *Advanced Language Processing*
- Applicazioni:
 - *Web Crawling, Web Search*
 - Classificazione ed annotazione semantica di testi, Musica ed Immagini

Organizzazione del Corso:

- *Parte I: Advanced Machine Learning*
 - ML, & Data Mining. Supervised ed Unsupervised Learning
 - Algoritmi generativi: Hidden Markov Modeling
 - Teoria Statistica dell'Apprendimento e Kernel-based learning.
 - Support Vector Machine. Structured Learning.
- *Sezione II: Statistical Language Processing*
 - Language Modeling and Sequence labeling
 - Statistical Parsing. Lexical Acquisition and Lexicalized Parsing.
 - Kernel-based NLP: SRL. Information Extraction.
- *Sezione III: Web Mining & Retrieval*
 - Sistemi e Tecnologie per l'IR nel Web.
 - Web Retrieval. Algoritmi di Ranking nel Web. Page Rank.
 - Introduzione al Social Network Analysis, Opinion Mining.

Obbiettivi (2):

Aspetti fondazionali

- Modelli ed Algoritmi di Apprendimento Automatico
- Ottimizzazione: Feature Selection, Dim. Reduction, Geometrical Embeddings
- *Statistical Learning Theory*
- Modelli di IR, motori di ricerca
- Esercitazioni
 - Completamento della Teoria
 - Introduzioni a sistemi e tecnologie
 - Progetto Finale
 - Applicazioni di modelli avanzati di IR
 - *ML for IR*
 - *Statistical Natural Language Processing*



Target

- Laurea Triennale in Informatica (Ing. Inf. Aut., Tel., Gest,)
- Prerequisiti:
 - ✓ Elementi di Analisi e Geometria
 - ✓ Elementi di Logica
 - ✓ Calcolo delle Probabilità e Statistica
 - ✓ Basi di Dati, Sistemi Basati su Conoscenza
- ✓ Vedi Corsi: ML, IR

Orari

- Lunedì: Aula C4 (Ed. Nuovo della Didattica, Sede Ingegneria), ore 9:30-11:15
- Martedì: Aula C4, ore 9:30-11:15
- Mercoledì: Aula C5 ore 9:30-11:15

Ricevimento: **MAR TEDI'** dopo la lezione

Testi Consigliati

- **IR** *Introduction to Information Retrieval*, Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze, Cambridge University Press, 2008. <http://nlp.stanford.edu/IR-book/>
- **MLxIR**. *Automatic Text Categorization: From Information Retrieval to Support Vector Learning*, R. Basili, A. Moschitti. ARACNE Ed., 2005
- **Materiale complementare**
 - Pattern Recognition and Data Mining, C. Bishop, 2006.
 - "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Liu, Bing, Springer-Verlag, Series: Data-Centric Systems and Applications, Berlin, ISBN: 978-3-540-37881-5, 2007.
 - Articoli e Dispense
 - Dispense e Lucidi delle lezioni:
<http://www.uniroma2.it/didattica/WmIR/>

<http://www.uniroma2.it/didattica/WmIR/>

Sito Web

- (per ora) Viene usato il sito Didattica Web di ateneo all'indirizzo:
 - http://www.uniroma2.it/didattica/Wm_IR

Organizzazione: Esami

- *Modalità del Corso*
 - *Full (9 crediti)*
 - *Informatica, Laurea Tecnologie di Internet*
 - *Light (6 crediti)*
 - *Laurea Ing. Informatica, Automatica e Gestionale*



Esami Light

- Esonero MidTerm (dopo il 20 Aprile)
- Scritto finale (dopo il 2 Giugno 2013)
- E' facoltativo (tirocinio?)
 - un progetto (ad esempio, una applicazione IR in ambiente mobile, Android)
oppure
 - un approfondimento bibliografico (ricerca bibliografica su un arg. avanzato)



Esami Full

- Esonero MidTerm (dopo il 20 Aprile)
- Scritto finale (dopo il 2 Giugno 2013)
- E' obbligatorio
 - un progetto (ad esempio, una applicazione IR in ambiente mobile Android)
 - oppure
 - un approfondimento bibliografico (ricerca bibliografica su un arg. avanzato)



Esami (I)

- **Approfondimento Bibliografico**
 - assegnamento di un tema e della relativa bibliografia,
 - ricevimento dedicato (fuori degli orari delle lezioni),
 - produzione di una relazione finale (stile Tesi),
 - Esame: presentazione finale della tesina (poss. entro la fine del Corso)

Esami (2)

- **Progetto** (max 2/3 persone)
 - Assegnamento di un progetto (stile tesi),
 - ricevimento dedicato (fuori degli orari delle lezioni),
 - produzione di un dimostrativo e di una breve relazione
 - Esame: Discussione del progetto + Demo



Esempi di Progetti

- Riconoscimento e tracking di opinion leaders in social networks
- Algoritmi graph-based per NLP
- Sviluppo di un sistema di classificazione su informazione musicale (audio files, Web pages, MIDI)
- Community Detection in Open Linked Data
- Titolazione automatica di immagini e video



Domande?



Action List

- Registrarsi al Corso :
 - URL:
<http://didattica.art.uniroma2.it/esami/>
- Definire i propri estremi e tipo di Corso (ad es. i CFU e o i Corsi già sostenuti) tramite il campo “Note”
- Verranno pubblicati:
 - Elenchi dei gruppi registrati
 - Progetti
 - Orari ricevimento per gli studenti che non seguono
 - Slide e materiali complementari (*in progress*)