Principali informazioni sull'insegnamento	
Titolo insegnamento	Accesso Intelligente all'Informazione ed Elaborazione del Linguaggio Naturale
Corso di studio	Magistrale in Informatica (LM18) – (corso di studio in disattivazione)
Crediti formativi	6
Denominazione inglese	Intelligent Information Access and Natural Language Processing
Obbligo di frequenza	No, ma la frequenza è fortemente consigliata
Lingua di erogazione	Italiano

Docente responsabile	Nome Cognome	Indirizzo Mail
	Giovanni Semeraro	giovanni.semeraro@uniba.it
Luogo ed Orario di Ricevimento	Ufficio 757 Dipartimento di Informatica Campus Universitario Via E. Orabona 4 70126 Bari	Martedì: 16:00-17:30 Giovedì: 16:00-17:30

Dettaglio credi formativi	Ambito disciplinare	SSD	Crediti
	Informatico	INF/01	6

Modalità di erogazione	
Periodo di erogazione	Primo semestre
Anno di corso	Secondo
Modalità di erogazione	Lezioni frontali Esercitazioni in aula

Organizzazione della didattica	
Ore totali	150
Ore di corso	62
Ore di studio individuale	88

Calendario	
Inizio attività didattiche	25/09/2017
Fine attività didattiche	12/01/2018

Syllabus	
Prerequisiti	Propedeuticità formali: nessuna Propedeuticità culturali: conoscenza di un linguaggio di programmazione (indispensabile), matematica discreta (importante), calcolo delle probabilità e statistica (importante), insegnamenti del primo anno di corso (importante)

Risultati di apprendimento previsti (declinare rispetto ai Descrittori di Dublino) (si raccomanda che siano coerenti con i risultati di apprendimento del CdS, compreso i risultati di apprendimento trasversali)	• *Conoscenza e capacità di comprensione* Il discente acquisirà la conoscenza dei fondamenti essenziali delle discipline dell'elaborazione del linguaggio naturale e dell'accesso intelligente all'informazione. In particolare, il discente sarà in grado di conoscere e saper comprendere: - gli aspetti teorici, metodologici e operativi dell'elaborazione del linguaggio naturale con particolare riferimento ai principali livelli di analisi linguistica; - le tecniche e le principali piattaforme open source per l'elaborazione del linguaggio naturale; - gli aspetti teorici, metodologici e operativi dell'accesso intelligente all'informazione e dell'indicizzazione semantica dell'informazione: - gli aspetti teorici, metodologici e operativi dei sistemi di filtraggio dell'informazione e dei recommender systems; - le tecniche e le principali piattaforme open source per la progettazione di recommender systems; • *Conoscenza e capacità di comprensione applicate* Il discente sarà in grado di: - applicare le conoscenze inerenti all'elaborazione del linguaggio naturale ed ai principali livelli di analisi linguistica per risolvere problemi di trattamento ed integrazione di dati non strutturati (testo) e di accesso intelligente all'informazione, anche in ambiti nuovi o non familiari o inseriti in contesti interdisciplinari; - utilizzare le tecniche e le principali piattaforme per l'elaborazione del linguaggio naturale e per la progettazione di recommender systems al fine di realizzare e valutare sistemi informatici complessi che richiedano il trattamento e l'integrazione di dati non strutturati (testo) ed il filtraggio dell'informazione; - applicare le proprie competenze sia per individuare soluzioni efficaci a problemi complessi inerenti al trattamento di dati non strutturati ed al filtraggio dell'informazione sia per giustificare, sostenere ed argomentare le proprie scelte. • *Autonomia di giudizio* Il discente sarà in grado di: - formulare una propria valutazione e definire un proprio giudizio critico e di sostenerlo nell'ambito del gruppo di lavoro con cui si è sviluppato il caso di

	studio; - integrare autonomamente le conoscenze e gestire la complessità derivante dalla limitatezza o incompletezza delle informazioni disponibili; - prendere decisioni ed individuare soluzioni nell'ambito dell'accesso intelligente all'informazione e dell'elaborazione del linguaggio naturale tenendo conto delle implicazioni sociali ed etiche e delle responsabilità professionali che esse comportano. • *Abilità comunicative* Il discente sarà in grado di: - scegliere la forma ed il mezzo di comunicazione adeguati agli interlocutori, sia specialisti sia non specialisti; - comunicare in maniera efficace informazioni, idee, problemi e soluzioni inerenti all'accesso intelligente all'informazione ed all'elaborazione del linguaggio naturale; • *Capacità di apprendere* Il discente acquisirà la capacità di: - sviluppare un alto livello di autonomia nell'apprendimento delle discipline dell'elaborazione del linguaggio naturale e dell'accesso intelligente all'informazione - tenersi aggiornato rispetto all'evoluzione delle discipline dell'elaborazione del linguaggio naturale e dell'accesso intelligente all'informazione attingendo a fonti bibliografiche sia in lingua italiana sia in lingua inglese - proseguire il proprio percorso formativo intraprendendo studi successivi ad alto grado di specializzazione.
Contenuti di insegnamento	1. Introduzione all'elaborazione del linguaggio naturale: Perché l'elaborazione del linguaggio naturale, hype cycle dell'elaborazione del linguaggio naturale, elaborazione del linguaggio naturale e linguistica computazionale, elementi di linguistica computazionale, panoramica sui livelli di analisi linguistica (fonetica, fonologia, morfologia, sintassi, semantica, pragmatica/logica), panoramica su analisi sintattica (categorie linguistiche, Part-of-Speech, categorie open e closed, collocations, shallow parsing, identificazione di strutture di base, full parsing, anaphora resolution), panoramica su analisi semantica (understand languages, textual entailment, textual similarity, word sense disambiguation, dictionaries,

	annotated examples, corpora), relazione tra elaborazione del linguaggio naturale e altre aree (information retrieval, learning semantics, information extraction, machine translation, personal digital assistants), il Loebner Prize. Ore lezione frontale: 2 2. Elementi di linguistica computazionale: Livelli di descrizione formale: fonetica, fonologia, morfologia, sintassi, semantica, pragmatica/logica. Categorie linguistiche: Part.of-Speech, categorie open e closed. Parole (words,), sintagmi (phrases), proposizioni (clauses), frasi/periodi (sentences). Ore lezione frontale: 3 3. Livello lessicale: Definizione di lessico o vocabolario, definizione di lessico attivo e lessico passivo, legge di Zipf. Normalizzazione. Tokenizzazione. Eliminazione delle stop word. Risorse per la tokenizzazione: espressioni regolari e tokenizzazione in Java, la classe File, leggere/scrivere file di testo, la classe String: espressioni regolari, esercizio "Leggere un file di testo e suddividerlo in token. Stampare un token per linea". Stemming: algoritmo di Porter. Risorse per lo stemming: stemming in Java, la libreria Snowball (metodi ed esempi). Lemmatizzazione. Risorse per la lemmatizzazione: Morph-it!, LemmaGen (Ripple Down Rules). Ore lezione frontale: 3 Ore esercitazione in aula: 8 4. Part.of-Speech (PoS) tagging: Definizione. Esempi. Applicazioni e motivazioni. Classi di parole. Insiemi di Tag. Esempi di tagging (Penn Treebank) di frasi del Brown Corpus. Il problema del PoS-tagging. Ambiguità nell'attribuzione delle classi di parole. Approcci al PoS-tagging: rule-based, stochastic (Hidden Markov Models), transformation-based (Brill tagger). Valutazione. Risorse per il PoS-tagging: I-CAB dataset, ACOPOST, Stanford PoS-tagger (demo). Ore lezione frontale: 3 Ore esercitazione in aula: 3 5. Livello sintattico: Parsing di grammatiche context-free (CFG). Top-down parsing. Bottom-up parsing. Parsing di grammatiche contect-free probabilistiche. Grammatiche a dipendenze. Ore lezione frontale: 3 6. Livello semantico: Semantica lessicale. Word Sense Disambiguation. Elementi di base (Word Sense Disambiguation e Word Sense Discrimination: definizioni, polisemia, ambiguità,). Metodi knowledge-based (Machine Readable Dictionaries, thesauri, reti

	semantiche, l'algoritmo di Lesk, somiglianza semantica, euristiche, selectional preferences). Metodi basati su grafi (page rank, esempi). Metodi supervisionati (Sense Tagged Text, classificatori singoli, ensemble di classificatori, bootstrapping: co-training, self-training, algoritmo di Yarowsky). Ore lezione frontale: 3 Ore esercitazione in aula: 3 7. Livello semantico: Word Sense Discrimination. Semantica distribuzionale (vettore distribuzionale, matrice parole-contesti, spazi geometrici, costruzione di uno spazio semantico). Random indexing (vettore contesto, lemma di Johnson-Lindenstrauss, esempi). Permutazioni. Modelli distribuzionali semantici (DSM) semplici. Modelli distribuzionali semantici (DSM) strutturati. Ore lezione frontale: 3 8. Piattaforme open source per l'elaborazione del linguaggio naturale: Apache OpenNLP, Stanford CoreNLP, FreeLing, LingPipe, UIMA, GATE. Esercizi. Ore esercitazione in aula: 3 9. Introduzione all'accesso intelligente all'informazione: Indicizzazione semantica utilizzando fonti esterne di conoscenza (WordNet, Wikipedia), indicizzazione semantica per accesso multilingua, Knowledge Infusion (KI) per la creazione di basi di conoscenza da fonti aperte, KI e giochi linguistici (cruciverba, OTTHO, la Ghigliottina). Esercizi. Ore lezione frontale: 2 Ore esercitazione in aula: 3 10. Open Data: Pubblica Amministrazione e Open Data, Open Government, obblighi normativi, linee guida e metodologie per rendere aperti i dati, classificazione dei dati aperti, licenze (le sei licenze Creative Commons, i quattro attributi BY, NC, ND, SA, l'Italian Open Data License), Linked Open Data cloud, il Web dei Dati. Esempi di creazione di specifiche RDF, esempi di microformati (RDFa), esempi di utilizzo dei vocabolari Dublin Core e Friend-Of-A-Friend, esempi di utilizzo del linguaggio di interrogazione SPARQL. Ore lezione frontale: 2 Ore esercitazione in aula: 1 11. Recommender Systems: Elementi di base. Decision making e information overload. Information Retrieval (IR) vs. Information Filtering: analogie, differenze. Definizione di recommender system, dominio del

	problema, pipeline. Principali paradigmi. Classificazione. Ore lezione frontale: 2 12. Collaborative Filtering (CF) e Content-Based Filtering (CB): User-to-User CF. Item-to-Item CF. Content-Based Filtering: rappresentazione dei contenuti, rappresentazione dei profili utente e calcolo della somiglianza tra contenuti e utenti. Semantics-aware Content-Based Recommender Systems. Ore lezione frontale: 2 Ore esercitazione in aula: 1 13. Altri paradigmi di Recommender Systems: Knowledge-based ecommender Systems, Hybrid Recommender Systems, Context-aware Recommender Systems. Ore lezione frontale: 1 14. Valutazione di Recommender Systems ed elementi avanzati: Valutazione in vitro. Valutazione in vivo. Disegno sperimentale. Metriche. Progettazione di user studies. Trasparenza e meccanismi di spiegazione. Novelty, diversity e serendipity. Cenni alla normativa EU General Data Protection Regulation (GDPR). Recommender Systems e social semantic web (social tagging, folksonomie, modellazione olistica degli utenti, social media e Linked Open Data). Ore lezione frontale: 3 15. Piattaforme open source per la progettazione di Recommender Systems: Apache Mahout. Introduzione. Casi d'uso. Algoritmi disponibili (clustering, topic modeling, classificazione, text processing, dimensionality reduction). Mahout e l'Apache Software Foundation. Mahout Samsara. Architettura generale. Componenti per la realizzazione di recommender systems. Componenti per la valutazione di recommender systems (metriche per la predizione, metriche basate su information retrieval). Come utilizzare Mahout. Ore esercitazione in aula: 8

Programma	
Testi di riferimento	1) D. Jurafsky and J. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentic-Hall Series in Artificial Intelligence, 2000. ISBN: 0130950696. 2) C. Manning and H. Schutze, Foundations of Statistical Natural Language Processing. MIT press, 2000. 3) Julia Hirschberg, Christopher D. Manning. Advances in natural language processing. Science, Vol. 349 no. 6245, pp. 261-266, 2015. DOI: 10.1126/science.aaa8685 3) Mark Stevenson, Word sense disambiguation: the case for

	combinations of knowledge sources. Stanford, CA: CSLI Publications, 2003. ISBN: 1575863898. 4) Eneko Agirre and Philip Edmonds, Word sense disambiguation: Algorithms and Applications. Springer Text, Speech and Language Technology, Vol. 33, 2007. ISBN: 1402048084. 5) Dominic Widdows, Geometry and Meaning. CSLI Publications, Stanford CA, 2004. ISBN: 1575864487. 6) Dietmar Jannach, Markus Zanker, Alexander Felfernig, Gerhard Friedrich, Recommender Systems: An Introduction. Cambridge University Press, 2010. ISBN: 0521493366. 7) F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook. 2nd Edition, Springer, 2015. ISBN: 9781489976369. 8) Armano, G.; de Gemmis, M.; Semeraro, G.; Vargiu, E. (Eds.), Intelligent Information Access. Studies in Computational Intelligence, vol.301, Springer, 2010. DOI: http://dx.doi.org/10.1007/978-3-642-14000-6. ISBN: 3642139994. 9) AA.VV. Open Data: come rendere aperti i dati delle PA - Linee guida per i siti web della PA. Vademecum. Formez PA, Ottobre 2011. 10) F. Di Donato, Lo stato trasparente: Linked open data e cittadinanza attiva. Edizioni ETS, 2011 ISBN: 9788846728876 (versione con licenza Creative Commons Attribuzione-Non commerciale 2.5 Italia: http://www.linkedopendata.it/wp-content/uploads/statotrasparente.pdf).
Note ai testi di riferimento	Le trasparenze mostrate a lezione, dispense integrative, letture consigliate e video utilizzati a lezione sono resi disponibili nella piattaforma di e-learning del Dipartimento di Informatica: http://informatica2.di.uniba.it
Metodi didattici	Lezioni frontali: 56 ore Esercitazioni in aula: 30 ore
Metodi di valutazione (indicare almeno la tipologia scritto, orale, altro)	Appelli d'esame (al termine dell'insegnamento) L'esame consta di una prova orale avente per oggetto la presentazione e la discussione di un caso di studio o progetto di natura sperimentale/applicativo, scelto tra quelli proposti durante l'insegnamento. La discussione della prova è individuale, mentre lo sviluppo del caso di studio o progetto può essere svolta in gruppi di massimo tre discenti. La durata della prova varia in base alla tipologia del caso di studio o progetto scelto, non superando comunque i 45 minuti.
Criteri di valutazione (per ogni risultato di apprendimento atteso su indicato, descrivere cosa ci si aspetta lo studente conosca o sia in grado di fare e a quale livello al fine di dimostrare che un risultato di apprendimento è stato raggiunto e a quale livello)	Nella prova orale il discente dovrà dimostrare di: - aver acquisito la conoscenza ed aver compreso approfonditamente gli aspetti teorici, metodologici e operativi propri dell'accesso intelligente all'informazione e dell'elaborazione del linguaggio naturale; - saper applicare in maniera appropriata le conoscenze

	inerenti all'elaborazione del linguaggio naturale ed all'accesso intelligente all'informazione per individuare soluzioni efficaci ai problemi incontrati nello sviluppo del caso di studio o progetto prescelto, anche integrando tecniche e componenti messe a disposizione dalle piattaforme open source oggetto di studio ed esercitazione; - saper giustificare adeguatamente le scelte di progetto effettuate, sostenendole attraverso argomentazioni critiche e dimostrando consapevolezza delle implicazioni sociali ed etiche oltre che delle responsabilità professionali che esse comportano; - saper comunicare in modo chiaro ed esauriente; - aver sviluppato un elevato livello di autonomia, anche attraverso l'enucleazione del proprio contributo nel caso di caso di studio o progetto sviluppato in gruppo; La valutazione della prova è espressa in trentesimi. La prova d'appello è superata con un minimo di 18/30. La determinazione del voto tiene conto dei seguenti aspetti: 1) correttezza delle soluzioni proposte nello sviluppo del caso di studio o progetto prescelto; 2) completezza delle soluzioni proposte nello sviluppo del caso di studio o progetto prescelto; 3) logica seguita dal discente nel proporre le soluzioni; 4) utilizzo di un adeguato formalismo per la descrizione delle soluzioni proposte nello sviluppo del caso di studio o progetto prescelto; 5) grado di innovatività delle soluzioni proposte nello sviluppo del caso di studio o progetto prescelto. Per superare la prova d'esame o la prova intermedia, il discente deve essere in grado di proporre una soluzione che soddisfi almeno l'aspetto 1). Voti superiori al minimo vengono attribuiti ai discenti in grado di sviluppare il caso di studio o il progetto prescelto in modo da soddisfare anche gli aspetti 2)-5).
Altro	