SAPIENZA
UNIVERSITÀ DI ROMA

Piano formativo

del Corso* di Formazione in:

Machine Learning, NLP e Web Scraping per l'analisi automatica di testi (II edizione)

Anno Accademico	2024-2025
Dipartimento	Scienze statistiche
Data Delibera approvazione di attivazione del corso in Dipartimento	19/03/2025
Direttore del Corso	Prof. Umberto Ferraro Petrillo
Numero minimo di ammessi	Tredici
Numero massimo di ammessi	Ventitré
Requisiti di ammissione	Diploma di scuola superiore. Per accedere al Corso si richiede il possesso di nozioni di base di Statistica ed una discreta conoscenza di un linguaggio di programmazione.
Obiettivi formativi	La prima parte del corso fornirà una breve introduzione al linguaggio di programmazione Python. Queste nozioni saranno utilizzate per affrontare compiti complessi legati allo scenario di analisi. Successivamente, verranno introdotti i principi di funzionamento dei siti web in generale e delle tecnologie utilizzate nella loro realizzazione e fruizione, quali i protocolli HTTP e REST, il linguaggio HTML, l'uso dei cookies. Infine, si introdurranno le tecnologie mediante le quali, da linguaggio Python, sarà

* Art. 1 punto 4 del Regolamento in Materia di Corsi di Master, Corsi di Alta Formazione, Corsi di Formazione, Corsi Intensivi D.R. 915/2018

- per Corso di Alta Formazione (CAF) il corso post - lauream professionalizzante di perfezionamento o approfondimento specialistico istituito in base alla L. 341/1990 art. 6. Vi si accede con la laurea, ha durata inferiore all'anno, consente l'acquisizione di massimo 20 Cfu e alla sua conclusione è rilasciato un attestato di frequenza;
- per Corso di Formazione (CF), il corso di aggiornamento professionale di durata inferiore all'anno che conferisce fino a un massimo di 10 Cfu. Vi si accede anche con il solo diploma di scuola media superiore e alla sua conclusione è rilasciato un attestato di frequenza;
- per Corsi Intensivi Summer/Winter School) i corsi, di norma residenziali, destinati a soggetti in possesso dei requisiti di cui all'art. 29 del presente regolamento, della durata da una a quattro settimane, connotati internazionalmente che conferiscono fino a un massimo di 10 Cfu e si concludono con il rilascio di un attestato di frequenza

	possibile, in maniera automatica e strutturata, acquisire i contenuti di pagine e siti web di interesse. Nella seconda parte del corso lo studente apprenderà le tecniche di analisi automatica del testo. Inizialmente verranno introdotti gli strumenti necessari per l'analisi, con particolare riferimento alle reti neurali profonde. Sarà introdotta la logica del Machine Learning e verranno illustrate le architetture di Neural Network più interessanti per l'analisi testuale. Si proseguirà quindi con le tecniche di pre-processing del testo. Si affronterà quindi il concetto di embedding di parole, frasi, documenti utilizzando l'approccio bag-of-words e la Latent Sematic Analysis. Successivamente si introdurranno i modelli Glove e Word2Vec e i pre-trained word vectors disponibili. A questo punto saranno introdotti i Transformers e i Large Language Models, che sono i modelli più recenti che hanno fatto fare un salto di qualità alle tecniche in questo ambito. In particolare, vedremo Bert e le sue varianti, con alcune applicazioni. Tratteremo quindi delle ultime proposte che comprendono anche i modelli più recenti e vedremo delle applicazioni con modelli open source.
Risultati di apprendimento attesi	- Padroneggiamento, da Python, delle principali tecniche di web scraping. - Analisi automatica di testi mediante le tecniche di NLP e machine learning viste a lezione.
Data di inizio delle lezioni	23/06/2025
Calendario didattico	○ 23 giugno 2025 ○ 24 giugno 2025 ○ 25 giugno 2025 ○ 30 giugno 2025 ○ 1° luglio 2025 ○ 2 luglio 2025

Stage	no
Modalità di erogazione della didattica	mista
CFU assegnati	4
Docenti Sapienza responsabili degli insegnamenti e relativi curricula brevi (max mezza pagina)	· Agostino Di Ciaccio Professore Universitario afferente al Dipartimento di Scienze Statistiche della SAPIENZA Università degli Studi di Roma; E' professore ordinario in Statistica presso l'Università di Roma La Sapienza e insegna "Data Mining e Classificazione", "Big Data Analytics", "Laboratory of Machine Learning"; è stato membro del consiglio direttivo della Società Italiana di Statistica e presidente del corso di Laurea in "Scienze statistiche e decisionali". Autore e referee di numerosi articoli scientifici. I suoi principali temi di ricerca riguardano: Data mining, Text and Web mining, machine learning, analysis of complex data, statistical inference. · Umberto Ferraro Petrillo Professore Universitario afferente al Dipartimento di Scienze Statistiche della SAPIENZA Università degli Studi di Roma; Umberto Ferraro Petrillo è attualmente Professore Ordinario in Informatica presso il Dipartimento di Scienze Statistiche dell'Università di Roma – 'La Sapienza', dove tiene gli insegnamenti di "Gestione ed Elaborazione di Big Data" e "Big Data Analytics" nell'ambito dei Corsi di Laurea Magistrale in "Statistica e Scienze Decisionali" e "Statistical Methods and Applications", nonché l'insegnamento di "Basi Dati", nell'ambito del Corso di Laurea in Statistica Gestionale. Si è laureato con lode in Scienze dell'Informazione nel 1997 presso l'Università degli studi di Salerno. Presso lo stesso Ateneo ha conseguito nel 2002 il Dottorato in Informatica. È stato titolare di un contratto di collaborazione con l'Istituto per le Tecnologie Industriali e l'Automazione del CNR. È stato inoltre titolare di assegno di ricerca all'Università degli studi di Salerno e all'Università degli studi di Roma – 'La Sapienza'. E' stato visiting fellow presso la School of Interactive Computing di Georgia Tech Institute da Ottobre 2016 a Febbraio 2017.

	I suoi temi di ricerca includono lo sviluppo e la sperimentazione di algoritmi distribuiti efficienti per la risoluzione di problemi di BioInformatica. Si occupa inoltre di problemi di Network Security e Computer Security, nonché di calcolo distribuito e supercalcolo.
Eventuali partner convenzionati	No
Sede di svolgimento Sapienza o sedi esterne (obbligo di Convenzione)	DIPARTIMENTO DI SCIENZE STATISTICHE
Quota di iscrizione prevista	900 QUOTA ORDINARIA in RATA UNICA
Eventuali quote di esenzioni parziali o totali dal pagamento della parte di quota di pertinenza del Dipartimento espresse in percentuali (numero intero) rispetto alla quota di iscrizione (max due tipi di esenzioni)	N. 7 QUOTE AGEVOLATE DI € 720 in RATA UNICA PER STUDENTI, ASSEGNISTI e DOTTORANDI. 10% di iscrizione sulla quota pro-capite, per gruppi di 3 o 4 persone iscritte da uno stesso Ente/Società, corrispondenti ad una quota individuale di 810 euro; 20% di iscrizione sulla quota pro-capite, per gruppi di 5 o più persone iscritte da uno stesso Ente/Società, corrispondenti ad una quota individuale di 720 euro.
Contatti di Segreteria	cristina.puteo@uniroma1it 06/49910502

SAPIENZA
UNIVERSITÀ DI ROMA

Piano delle Attività Formative

(Insegnamenti, Seminari di studio e di ricerca, Stage, Prova finale)

Denominazione attività formativa	Responsabile insegnamento	Settore scientifico disciplinare	CFU	Ore	Tipologia	Lingua
Attività I: Introduzione a Python	Prof. Umberto Ferraro Petrillo	INF/01	1	10	Lezione frontale	Italiana
Attività II: Web Scraping	Prof. Umberto Ferraro Petrillo	INF/01	1	10	Lezione frontale	Italiana
Attività III: Machine Learning	Prof. Agostino Di Ciaccio	SECS/01	1	10	Lezione frontale	Italiana
Attività IV: NLP	Prof. Agostino Di Ciaccio	SECS/01	1	10	Lezione Frontale	Italiana

Prova finale	Esercitazione	SSD non previsto			*Prova di laboratorio*
Altre attività		SSD non previsto			
		TOTALE CFU	**4**		

Il numero minimo di Cfu assegnabili ad una attività è 1 (ai sensi dell'art. 23 del Regolamento didattico d'Ateneo si precisa che 1 CFU corrisponde 6 – 10 ore di lezione frontale, oppure 9 - 12 ore di laboratorio o esercitazione guidata, oppure 20 - 25 ore di formazione professionalizzante a piccoli gruppi o di studio assistito).