Corso introduttivo:
Data Science[1] con Python

Emmanuele Somma
(Servizio Struttura Economica — Divisione Biblioteca)

La comunità internazionale dei Data Scientists, in special modo quelli coinvolti nei progetti di Big Data, ha scelto Python come linguaggio professionale per l'analisi esplorativa dei dati e per la realizzazione di applicazioni e elaborazioni scientifiche. Per usare l'espressione dell'astrofisico dell'Università di Berkley Josh Bloom, Python è diventato la «Super-Colla per un moderno Workflow Scientifico» per la ricerca riproducibile. Python è apprezzato per la vastità delle funzioni contenute nei suoi moduli standard e per la potenza delle librerie scientifiche (NumPy per il calcolo numerico, SciPy per gli algoritmi, pandas per le serie storiche, matplotlib per la visualizzazione e IPython per l'interazione). Conoscere Python è, allo stato attuale, irrinunciabile per poter confrontarsi con il mondo della Data Science, Big Data e Machine (o Deep) Learning.

Python è un linguaggio general-purpose con una sintassi minima, semplice e intuitivo ed ha quindi una curva di apprendimento relativamente piatta soprattutto per chi ha già la conoscenza di qualche altro linguaggio di programmazione scientifico, come Matlab, Stata o R.

Nella prima parte il corso introduce il linguaggio di programmazione con una particolare attenzione all'ambito di riferimento dell'analisi dei dati, l'applicazione di algoritmi scientifici, e la visualizzazione dei dati.

1 La Data Science, nota anche come data-driven science, è un campo interdisciplinare basato su metodi, processi e sistemi scientifici per estrarre conoscenze o insight dai dati in varie forme, strutturate o non strutturate, simili al data mining. La Data Science mira ad unificare la statistica, l'analisi dei dati e la computer science e relativi metodi per comprendere e analizzare i fenomeni reali di cui sono disponibili i dati. Impiega tecniche e teorie tratte da molti campi all'interno delle vaste aree della matematica, delle statistiche, delle scienze dell'informazione e dell'informatica, in particolare dai sottodomini di apprendimento automatico, classificazione, analisi dei cluster, data mining, database e visualizzazione.
Il vincitore del premio Turing, Jim Gray, ha immaginato la data science come un "quarto paradigma" della scienza (empirico, teorico, computazionale e ora basato sui dati) e ha affermato che "tutto sulla scienza sta cambiando a causa dell'impatto della tecnologia dell'informazione" e della sempre maggiore quantità di dati grezzi a disposizione.

Nella seconda parte saranno invece proposte lezioni monografiche su alcuni temi di particolare interesse come l'analisi dei dati testuali e provenienti dai social media,l'acquisizione dei dati da web (web-scraping), l'analisi predittiva e alcune tecniche di machine-learning, nonché di accesso ai dati del semantic-web.

Prerequisiti:

Il corso è rivolto a ricercatori, tecnologi o assistenti di ricerca con una pregressa conoscenza o di un linguaggio di programmazione generico o scientifico, o di un pacchetto scientifico per l'analisi dei dati.

Non c'è bisogno comunque che la conoscenza di un linguaggio di programmazione sia estesa ma si assume che si conosca quantomeno:

- come creare, assegnare e usare le variabili
- come scrivere programmi con i cicli (loop)
- come scrivere programmi con scelte condizionali (if)
- come realizzare e usare e le funzioni

Per poter seguire gli esempi durante il corso è utile portare dietro un computer portatile con una recente installazione della distribuzione scientifica Anaconda Python.

Obiettivi di apprendimento:

- Sintassi, strutture dati e controllo del linguaggio Python
- Procedure basilari della data-science
- Utilizzo degli ambienti di lavoro Python e Jupyter
- Comprensione delle tecniche di manipolazione di dataset
- Analisi scientifica di base e principali metodi di machine-learning
- Rappresentare in modo efficace i risultati in forma grafica

Calendario (parziale e provvisorio):

Gennaio 2018:
- Mercoledì 10 (2 ore)
- Martedì 16
- Mercoledì 24
- Mercoledì 31

Febbraio 2018:
- Mercoledì 7
- ...

Sede: *Centro Carlo Azeglio Ciampi per l'educazione monetaria e finanziaria*
Via Nazionale 190 – 00184 Roma

Programma:

- Caratteristiche base del linguaggio
- Caratteristiche avanzate
- Ambienti di lavoro: Python, IPython, Spyder e Jupyter
- Visualizzazione dei dati
- Plotting interattivo e Interfacce utente grafiche
- Customizzare IPython
- Calcolo Numerico con NumPy
- Algoritmi scientifici con SciPy
- Data mining con Scipy
- Analisi delle serie storiche (pandas)
- Registrare, processare e conservare i dati
- Lavorare con i database
- Web-scraping dei dati su Internet
- Calcolo High-performance e Parallelo
- Analizzare i dati testuali e i social media
- Analisi predittiva e machine learning
- Il Web semantico e i linked open data