Big Data Tools

Irene Siragusa, PhD

Università
degli Studi
di Palermo

Outline

- General information
- Pre-requisites
- Course objectives
- Topics
- Textbooks
- Tools

General information

Big Data (12 ECTS)

Big Data Tools

- 1st semester
- 6 ECTS course

Big Data Analytics

- 2nd semester
- 6 ECTS course

General information

Lectures, where and when?

- Tuesday 08:00 – 10:00 @ Aula Informatica ex DSSM building 13
- Thursday 15:00 – 17:00 @ Aula Informatica ex DSSM building 13

Contact me?

- irene.siragusa02@unipa.it w/ subject [BIG-DATA]

General information

?*Course material*

- BIG-DATA-TOOLS GitHub https://github.com/iresiragusa/BIG-DATA-TOOLS/
  - Slides -> theory slides
  - Exercises -> python notebooks and related material

?Subscribe from the university website to the course for official communications

Pre-requisites

- Basic knowledge of probability, statistics, and linear algebra
- Basic Python programming
- Relational DB, basic SQL

Course objectives

- The course is aimed at providing students with a knowledge of the software architectures for Big Data along with both the main algorithms for data analysis and preprocessing techniques with the aim of developing autonomously whole data analysis pipelines for real case studies.
- This course is arranged in lessons and exercise sections.
- Any doubts? Just ask (I'll do my best in answering)

Course objectives - Theory

- Data analysis process
  - Pre-processing techniques
  - Similarity measures in data analysis
  - Frequent patterns analysis algorithms
- Software architectures for Big Data
  - Databases noSQL
  - MapReduce algorithm
  - Apache Hadoop ecosystem
  - Apache Spark framework

Course objectives - Practical

- Data analysis process
  - Pre-processing techniques
  - Similarity measures in data analysis
  - Frequent patterns analysis algorithms
- Software architectures for Big Data
  - Databases noSQL
  - MapReduce algorithm
  - Apache Hadoop ecosystem
  - Apache Spark framework

Topics

- Data analysis workflow
  - data gathering, preprocessing, applying analysis techniques, knowledge extraction.
- Data preparation
  - data types, data cleaning, missing data, sampling.
- Dimensionality reduction
  - Principal Component Analysis, Singular Value Decomposition, Wavelet transform, Muti Dimensional Scaling, Graph embedding.
- Similarities and distances for different data types
  - quantitative data, categorical data, text data, temporal sequences, graphs.

Topics

- Mining frequent patterns
  - Apriori algorithm, correlation statistics.
- Software architectures for Big Data
  - database noSQL, MongoDB, Data lake.
- Software architectures for Big Data
  - MapReduce algorithm, Apache Hadoop, HDFS.
- Software architectures for Big Data
  - Apache Spark and its libraries.

Textbooks

- Data Mining: The Textbook, 2015, Charu C. Aggarwal, Springer-Verlag New York, ISBN 978- 3319141411, https://link.springer.com/book/10.1007/978-3-319-14142-8
- Spark: The Definitive Guide: Big Data Processing Made Simple, 2018, di Bill Chambers e Matei Zaharia, Oreilly & Associates Inc, ISBN 978-1491912218
- Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and The Cloud, 2019, di Paul Deitel & Harvey Deitel, Pearson, ISBN 978-0135404676

Tools

- VSCode
  - https://code.visualstudio.com/download

- Python ≥ 3.7
  - [W10-W11] Microsoft store
  - [Linux] packet manager (apt, snap, …)
  - [MacOS] brew
  - [all] https://www.python.org/downloads/