COM-490

Large-scale data science for real-world data

Bouillet Eric Pierre, Delgado Pamela, Sarni Sofiane, Verscheure Olivier

Cursus	Sem.	Type
Data Science	MA2, MA4	Opt.
Data science minor	E	Opt.
Electrical and Electronical Engineering	MA2, MA4	Opt.

Language of teaching	English
Credits	6
Withdrawal	Unauthorized
Session	Summer
Semester	Spring
Exam	During the semester
Workload	180h
Weeks	14
Hours	**4 weekly**
Practical work	4 weekly
Number of positions	

It is not allowed to withdraw from this subject after the registration deadline.

Summary

This hands-on course teaches the tools & methods used by data scientists, from researching solutions to scaling up prototypes to Spark clusters. It exposes the students to the entire data science pipeline, from data acquisition to extracting valuable insights applied to real-world problems.

Content

1. Crash-course in Python for data scientists

- Main Python libraries for data scientists
- Interactive data science with web-based notebooks
- Reusable compute environments for reproducible science
- Homework: Curating data from a network of CO2 sensors

2. Distributed data wrangling at scale

- Understand the main constituents of an Apache Hadoop distribution
- Put Map-Reduce into practice
- Focus on HDFS, Hive and HBase and associated data storage formats
- Homework: Big data wrangling with massive travel data from SBB/CFF

3. Distributed processing with Apache Spark

- RDDs and best practices for order of operations, data partitioning, caching
- Data science packages in Spark: GraphX, MLlib, etc.
- Homework: Uncovering world events using Twitter hashtags

4. Real-time big data processing using Apache Spark Streaming

- Window-based processing of unbounded data
- Homework: Geospatial analysis and visualization of real-time train geolocation data from the Netherlands

5. Final project - Summing it all up

- Robust Journey Planning on the Swiss multimodal transportation network - Given a desired departure, or arrival time,your route planner will compute the fastest route between two stops within a provided uncertainty tolerance

expressedas interquartiles. For instance, Ã¢##what route from A to B is the fastest at least Q% of the time if I want to leave from A(resp. arrive at B) at instant t?Ã¢##.

Keywords

Data Science, IoT, Machine Learning, Predictive Modeling, Big Data, Stream Processing, Apache Spark, Hadoop, Large-Scale Data Analysis

Learning Prerequisites

Required courses

Students must have prior experience with Python

Recommended courses

Students must have prior experience with at least one general-purpose programming language.

Important concepts to start the course

It is recommended that students familiarize themselves with concepts in statistics and standard methods in machine learning.

Learning Outcomes

By the end of the course, the student must be able to:

- Use standard Big Data tools and Data Science librairies
- Carry out out real-world projects with a variety of real datasets, both at rest and in motion
- Design large scale data science and engineering problems
- Present tangible solution to a real-world Data Science problem

Transversal skills

- Demonstrate a capacity for creativity.
- Plan and carry out activities in a way which makes optimal use of available time and other resources.
- Write a scientific or technical report.

Teaching methods

- Hands-on lab sessions
- Homework assignments
- Final project

... using real-world datasets and Cloud Compute & Storage Services

Expected student activities

- STUDY : Attend the lab sessions
- WORK : Complete homework assignments
- ENGAGE : Contribute to the ineractive nature of the class
- COLLABORATE : Work in small groups to provide solutions to real-world problems

- EXPLAIN : Present ideas and results to the class

Assessment methods

- 60% continuous assessment during the semester
- 40% final project, done is small groups

Supervision

Office hours	Yes
Assistants	Yes
Forum	Yes

Resources

Virtual desktop infrastructure (VDI)

No

Bibliography

- Python Data Science Handbook: Essential Tools for Working with Data by Jake VanderPlas, O'Reilly Media, November 2016
- pyGAM - https://github.com/dswah/pyGAM

A list of additional readings will be distributed at the beginning of the course

Websites

- https://dslab2020.github.io

Moodle Link

- https://go.epfl.ch/COM-490