Data acquisition, extraction and storage

ECTS : 4

Volume horaire : 24

Description du contenu de l'enseignement :

The objective of this course is to present the principles and techniques used to acquire, extract, integrate, clean, preprocess, store, and query datasets, that may then be used as input data to train various artificial intelligence models. The course will consist on a mix of lectures and practical sessions. We will cover the following aspects:

- Web data acquisition (Web crawling, Web APIs, open data, legal issues)
- Information extraction from semi-structured data
- Data cleaning and data deduplication
- Data formats and data models
- Storing and processing data in databases, in main memory, or in plain files
- Introduction to large-scale data processing with MapReduce and Spark
- Introduction to the management of uncertain data

Compétence à acquérir :

Understanding:

- how to acquire data from a variety of sources and in a variety of formats
- how to extract structured data from unstructured or semi-structured data
- how to format, integrate, clean data sets
- how to store and access data sets

Mode de contrôle des connaissances :

Project (50% of the grade) and in-class written assessment (50% of the grade)