[Data Mining for Big Data]

Course level: Master ([M2]) **Track(s):** [MLDM, DSC]

ECTS Credits: 5

Course instructors: [Elisa Fromont, Christine Largeron]

Education period: [3rd] semester **Language of instruction:** English

Expected prior-knowledge: [Data Analysis, Data Mining and Knowledge Discovery]

Aim and learning outcomes: This course focuses on data mining of very large data (i.e. data that do not fit in main memory) and on a number of algorithms specifically designed to be used on massive data (such as data extracted from the web, e.g. social networks, advertisement, recommender systems, …). It explains the principle of distributed file systems and shows Map reduce as a tool for creating parallel algorithms.

Keywords: [Big Data, Network Analysis, Text and Web Mining, Recommendation Systems, Map Reduce, Hadoop, Spark, LSH, R]

Syllabus:

- MapReduce, Hadoop/Spark and how to scale the usual data mining methods to big data (clustering, PCA, SVM): 9h

- Finding Similar Items in big data (LSH, KNN): 3h

- Mining Social-Networks Data: 6h

  - From social network to information network
  - Networks representation, visualization
  - Network analysis: measures and metrics
  - Models of network generation
  - Community detection (percolation algorithms, cliques, Mincut, spectral clustering)
  - Influence/ link prediction

- Text Mining: 6h

  - Overview of text mining

- Text preprocessing
- Features extraction - indexing
- Weighting models
- Document similarity
- Features Extraction - dimension reduction
- Text mining (categorization – clustering- association)
- Topics models (LSA, PLSA, LDA)

- Advertising the web: 3h

- Recommendation systems: 3h

- Practical sessions on Hadoop, Spark and R:10h

Organisation and timetable: [Volume CM/TD/TP] Lectures (15h), tutorials (15h) and lab sessions (10h).

Form(s) of Assessment: 1 theoretical exam (2h, coeff 2/3), 1 project: case study on a data mining challenge (coeff 1/3)

Literature and study materials:

- Mining of Massive Datasets (http://www.mmds.org/): Jure Leskovec, Stanford University, California, Anand Rajaraman, Milliways Laboratories, Jeffrey David Ullman, Stanford University, California, 2014.

- Data mining concepts and techniques (J. Han, M. Kamber, J. Pei)

- Mastering Text Mining with R6 de Kumar Ashish et Avinash Paul A. (April 2016)

Additional information/Contacts:

elisa.fromont@univ-st-etienne.fr

christine.largeron@univ-st-etienne.fr