Course Overview

Processamento de Streams

https://smduarte.github.io/ps2023/

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

Organization

- **Distributed Stream Processing Systems**
  - Sérgio Duarte (smd@fct.unl.pt)
  - 1/2 of the lectures (6 weeks)

- **Data Processing and Machine Learning for Streams**
  - Cláudia Soares (claudia.soares@fct.unl.pt)
  - 1/2 of the lectures (6 weeks)

Goal

- **Learn the fundamental concepts, languages, and systems for building applications that process data streams. This course discusses, presents generalist systems for real-time stream processing, and will focus on the study of systems for structured-data flow-oriented models.**
- **Knowledge**
  - Know the main programming models for streaming data processing
  - Know the languages and assimilate the fundamental characteristics to solve problems in the stream processing domain.
  - Understand the advantages and disadvantages of stream processing platforms.
- **Know how**
  - Being able to choose the most appropriate models, languages and tools to solve a stream processing problem.
  - Be capable of developing and executing stream processing applications using current tools and technologies.

Syllabus

- **Distributed Stream Processing Systems.**
  - System models for stream processing: streams as sequences of mini-batches (e.g. Spark streaming); continuous processing (e.g. Apache Flink, Storm).
  - Programming models.
  - System aspects: distribution, scalability and fault-tolerance.
  - Distributed time-series databases.
  - Systems for IoT stream processing.

- **Machine Learning for Streams.**
  - Introduction to learning from data
  - Dimensionality reduction for streams.
  - Learning under concept drift.
  - Incremental learning
  - Learning under imbalance / Learning from graphs.

Assessment

- **2 midterms [25%+25%]**
  - Late April and Mid June.
- **2 projects [25%+25%]**
  - Groups of 2 **students at most**
  - Project 1: **Late April**
  - Project 2: **Mid June**
- **Requirements to succeed**
  - Average of tests ≥ 8.0 ; Average of projects ≥ 10.0
- **Exam**: replaces one or both midterms.

Planning

Classes	Lectures	Labs	Obs.
6-mar.	Intro to big data frameworks		
13-mar.	Non-structured programming	Spark streaming	
20-mar.	Structured programming and SQL	Spark streaming SQL	
27-mar.	Continuous streaming	Kafka	
3-Apr	Stream processing ecosystem	TBD	
10-abr.	**(Easter)**	Student support	
17-Apr	Storage for streamable data	Student support	
24-abr.	Intro to learning from data	Refresher on Algebra and probability	
5-mai.	Dimensionality reduction	Intro to learning.	Friday
8-May	Learning under concept drift	Dim redux.	
15-May	Incremental learning	Concept drift.	
22-May	Imbalance in data streams	Incremental ML	
29-May	Learning from graphs	Data imbalance.	
5-jun.	Revisions	Graphs.	
12-jun.			

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

Bibliography

- **Processing Flows of Information: From Data Stream to Complex Event Processing**, GIANPAOLO CUGOLA and ALESSANDRO MARGARA, Politecnico di Milano, ACM Computing Surveys, Vol. 44, No. 3, Article 15, Publication date: June 2012
- **Event Processing in Action**, OPHER ETZION PETER NIBLETT, 2011, Manning Publications Co
- **Data Stream Management,** Lukasz Golab and Tamer Özsu. Morgan and Claypool, 2010.

Bifet et al., (2018) Machine Learning for Data Streams, MIT Press

This book does not cover all the material of the ML4Streams part, and we will not cover all topics in it. We will provide surveys and other papers as we go.

- **Papers will be provided during the semester**

NOVA SCHOOL OF SCIENCE & TECHNOLOGY