3 ECTS

14D008

Topics in Big Data Analytics I

Overview and Objectives

Constant advances in digital sensors, Internet, mobility and storage, result in the explosion of available data that potentially carries significant value to business, science and society. This poses many challenges both technological and analytical. A wide variety of techniques have arisen with the objective of discovering hidden patterns in data. These methods are fully exploited by top technology companies such as Amazon, Netflix, Twitter or Google and define the core of their competitive advantage. This course is structured as a series of four lectures where students will be presented with the theoretical underpinning and practical implementation of some of the most groundbreaking big data applications.

Lecturers

The course is coordinated by Emiliano Carluccio (Kernel Analytics) and classes will also be delivered by Mikel Arizaleta, Marçal Molins, Diego Gruber and Miquel Camprodon (Kernel Analytics).

Course Outline

1. **Hadoop administration for developers.** The day in which everyone should be able to create and use a standard Hadoop cluster is already here. We propose a 4 hour course with the basic tips & tricks of Hadoop administration that every data scientist should now.

2. **Deploying ML Models as APIs:** Web services have become ubiquitous in data science, both as a way to access data as well as an interface for models that simplify integration within an organization's systems. We will learn the building blocks of REST APIs, how to use them to retrieve data as well as how to deploy your own model as an API using R and OpenCPU.

3**. Personalized recommendation on distributed architectures.** An update of last year module making it really "big data". How to prepare data and execute collaborative filter algorithms on big data architectures (Hadoop, Spark)

4. **How to work with spatial data:** Maps, gps localizations, addresses, the spatial data of an entity give us a lot of new hidden information about the entity itself and about its neighbours. In this 4 hours class we will show how to get, clean and cross this data and how to extract the relevant information out of it.

5. **Dockers for analytics + practical modelling with spark(&)R**. In this 4 hours class we will show how to leverage Spark, R & Python to scale basic models to match whatever data volume. At the same time we will introduce Dockers as a mainstream alternative for software standardization, deployment and fast prototyping in analytics.

Required Activities

Attendance at classes, a practical project (consisting of programming exercises)

Evaluation

• 100% Project. Students will hand out a programming project for each topic at the end of the course.

14D008

3 ECTS

Topics in Big Data Analytics I

Materials

Books:

Anand Rajaraman and Jeffrey David Ullman. 2011. Mining of Massive Datasets. Cambridge University Press, New York, NY, USA. Chapters 3, 8 and 9.

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press.

Hadoop Operations, Eric Sammer, O'Reilly
Hadoop, the Definitive Guide, Tom White, O'Reilly

Agile Data Science: Building Data Analytics Applications with Hadoop, Russell Jurney, O'Reilly

Others:
Bennett, James, and Stan Lanning. "The netflix prize." Proceedings of KDD cup and workshop. Vol. 2007. 2007.

Bell, Robert M., and Yehuda Koren, "Lessons from the Netflix prize challenge."ACM SIGKDD Explorations Newsletter 9.2 (2007): 75-79.

Coordinate Sysems: Coordinate Systems and Map Projections, D.H. Maling
Carto: https://carto.com/docs/
^

MapReduce: Simplified Data Processing on Large Clusters

An Introduction to APIs by Brian Cooksey (Zapier)

Out of the Weeds and into Product: APIs and the Future of Data Science by Ken Tien (Mulesoft)

R Packages by Hadley Wickham (RStudio)

The OpenCPU System: Towards a Universal Interface for Scientific Computing through Separation of Concerns by Jeroen Ooms (OpenCPU)