

14D008

3 ECTS

Topics in Big Data Analytics

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 Pau Agulló (Kernel Analytics) and classes will be delivered by Mikel Arizaleta, Emiliano Carluccio, Marçal Molins, Roger Forcada (Kernel Analytics)

Course Outline

Google page rank algorithm

- Algebraic foundation and relevant theorems
- The Google page rank algorithm: definition and basic resolution
- Relevant R packages

Real time bidding: online advertisement

- Online advertising as an auction problem
- Approaches to the matching problem
- Formalization of the AdWords problem and real world implementation

Recommendation systems: the Netflix algorithm

- Collaborative Filtering
- User-User k-Nearest Neighbor Approach
- Hybrid Algorithms: evaluation and error metrics

14D008

3 ECTS

Topics in Big Data Analytics

Text analytics: Sentiment analysis on Twitter

- Data preparation: Parsing, Tokenization, Stemming, Regular expressions
- Document classification and measuring document similarity
- Resources for text mining: Annotated corpora, Text mining packages

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.

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.

Segaran, Toby. *Programming collective
intelligence: building smart web 2.0
applications*. " O'Reilly Media, Inc.", 2007.
Chapter 2.

Sholom M. Weiss, Nitin Indurkha, Tong
Zhang, *Fundamentals of Predictive Text
Mining*, Ed. Springer.

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

Others:

Aranyak Mehta, Online Matching and Ad
Allocation, Foundations and Trends in
Theoretical Computer Science, vol. 8 (4)
(2013), pp. 265-368.

S. Brin, L. Page, The anatomy of a large-scale
hypertextual web search engine, Computer
Networks and ISDN Systems, 30 (1-7) (1998),
pp. 107-117.

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.

Aranyak Mehta, Amin Saberi, Umesh Vazirani,
and Vijay Vazirani. Adwords and generalized
online matching. J. ACM, 54(5):22, 2007.