CAIM: Cerca i Anàlisi d’Informació Massiva
FIB, Grau en Enginyeria Informàtica

Slides by Marta Arias, José Luis Balcázar, Ramon Ferrer-i-Cancho, Ricard Gavaldá
Department of Computer Science, UPC

Fall 2018
http://www.cs.upc.edu/~caim
0. Presentation
Instructors

- Marta Arias (lectures + exercises)
  - marias@cs.upc.edu
  - Omega 322, 93 413 7858

- José Luis Balcázar (lab viernes)
  - jose.luis.balcazar@upc.edu
  - Omega 255, 93 413 7847

- Javier Béjar (lab martes y viernes)
  - bejar@cs.upc.edu
  - Omega 204, 93 413 7879
Class Logistics

- Thursdays, 14–16, A6202
  - Theory and exercises. Often, exercises will be proposed in advance.

- Fridays and Tuesdays, lab sessions
  - Fridays (group 11: 16–18, A5S102; group 13: 18–20, A5S109); Tuesdays 14–16 (group 12, C6S302).
  - Guided lab activities; expected to be complemented with an average estimate of 2 additional hours per session of autonomous work.
  - Some lab sessions will finish by handing in a short written report; these count towards the evaluation of the course.
Lab work - important rules

- Lab is done in pairs. Exceptions must have *prior* permission
- At most one assignment by the same two people - change partners each time
- Do not exchange information with others, other than general ideas; that will be considered plagiarism
Exercises

► In class, we will solve only a part of the exercises proposed
► You are strongly encouraged to try and solve the rest of the exercises
► Self-study: One or more small topics will not be explained in class. They will appear in the exam.
Evaluation

- Evaluation: as per “Guia Docent”


- On the day of Parcial 2 you may choose to do instead a final exam (F) on the whole course

- 40% Lab + max(30% P1 + 30% P2, 60% F)
First half (until midterm):

- Core Information Retrieval:
  - Introduction: Concept. The IR process
  - Information Retrieval Models
  - Indexing and Searching, Implementation
  - Information Retrieval Evaluation, Feedback Models

- Web Search:
  - Link analysis: Page Rank
  - Crawling the web
  - Architecture of a Web search system
Second half:

- The “Big Data” Slogan
  - Architecture of large-scale web search systems
  - The Map-Reduce paradigm
  - Introduction to NoSQL databases
  - The Apache ecosystem for web search.

- Social Network Analysis:
  - Characterizing of real complex networks
  - Communities, influence, information diffusion

- Clustering and Locality Sensitive Hashing

- Recommender Systems


Russell, Matthew, Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Site. O’Reilly, 2011

... There’s a whole web out there