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

Slides by Marta Arias, José Balcázar, Ricard Gavaldá
Department of Computer Science, UPC

Fall 2017
http://www.cs.upc.edu/~caim
0. Presentation
Instructors

- Ramon Ferrer-i-Cancho (lectures + exercises)
  - rferrericancho@cs.upc.edu
  - Omega S124, 93 413 4028

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

- Mondays, 12–14, A4102
  - Theory and exercises. Often, exercises will be proposed in advance.

- Thursdays, 8–10 (group 11, C6S306; group 13, A5S104); Fridays 12–14 (group 12, C6S308).
  - 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”
- Partial1: Nov ?, Partial2: Jan 18
- On the day of Partial2 you may choose to do instead a final exam on the whole course
- 40 % Lab + max(30 % Partial1 + 30 % Partial2, 60 % Final)
Contents I

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:

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

- Clustering and Locality Sensitive Hashing

- 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.

- Recommender Systems
References

- 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