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 2023

`http://www.cs.upc.edu/~caim`

0. Presentation

Instructors

- Ramon Ferrer-i-Cancho (lectures + exercices 10)
  - `rferrericancho@cs.upc.edu`
  - Omega 220, 93 413 4028
- Ignasi Gómez (lab 13)
  - `ignasi.gomez@upc.edu`
- Albert Calvo (labs 11 & 12)
  - `albert.calvo.ibanez@upc.edu`

Class Logistics

- Mondays, 12–14.
  - Theory and exercises. Often, exercises will be proposed in advance.
- Lab sessions: Thursdays and Fridays.
  - Guided lab activities; expected to be complemented with an average estimate of 2 additional hours per session of autonomous work.
  - 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
- 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”
- Parcial 1 (P1): during the Week of exams (usually late October or early November), Parcial 2 (P2): January. Check exact date & time of P1 and P2 here: `https://www.fib.upc.edu/ca/estudis/graus/grau-en-enginyeria-informatica/examens`
- 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)

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

Contents II

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

Bibliography

- R. Baeza-Yates, B. Ribeiro-Neto: Modern Information Retrieval (2nd ed.). Addison Wesley, 2010.
- I.H. Witten, A. Moffat, T. Bell: Managing Gigabytes. Morgan Kaufmann, 1999.
- C.D. Manning, P. Raghavan, H. Schütze: Introduction to Information Retrieval. Cambridge 2008.
- Z. Markov, D.T. Larose: Data Mining the Web. Wiley, 2007.
- 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