UNIVERSITÀ DI PISA

**DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE**

**Dottorato di Ricerca in Ingegneria dell'Informazione**

Doctoral Course

# "Challenges in Modern Web Search"

Prof. Franco Maria Nardini, Salvatore Trani
*ISTI-CNR - Italy*
*{name.surname}@isti.cnr.it*

**Short Abstract:** This PhD course focuses on Web search and discusses the challenges in the three main areas of Web search: i) crawling, ii) indexing, and iii) query processing. The course introduces each area by discussing the state of the art in the field and by presenting the open research questions. The emphasis of the course is on query processing, an area where machine learning provides an important contribution to advance the state of art. After an introduction of the different query processing techniques, the course i) introduces supervised techniques explicitly focused to target the ranking problem, ii) discusses several efficiency/effectiveness trade-offs in query processing and iii) analyse several related optimization techniques. The course will also provide an overview of the query processing techniques employing deep neural networks. Two hands-on sessions will cover indexing and query processing of public Web collections.

**Course Contents in brief:**

- Modern Web Search (**4 hours**)
    - The web: history, peculiarities and the importance of the search.
    - Anatomy of a modern Web search engine: crawling, indexing, query processing.
    - Crawling: definition and application. Architecture of a modern crawler.
    - Challenges in crawling the Web
- Fast Indexes for Web search (**4 hours**)
    - Data structures for indexing Web documents
    - Modern techniques for efficient text retrieval
    - Data structures for efficient k-NN search and retrieval over learned representations
    - Challenges in indexing the Web
    - Hands On: Indexing and basic query processing on a public Web collection
- Machine learning in modern query processors (**8 hours**)
    - Machine learning approaches for IR: Learning to Rank
    - Efficiency/Effectiveness Trade-offs, Cascading Architectures
    - Neural information retrieval and the role of pre-trained large language models
    - Dense/Sparse retrieval
    - Hands On: Learning to Rank and Deep Neural Networks for efficient Web search

**Total # of hours of lecture**: 16

**CV of the Teachers**

Franco Maria Nardini (http://hpc.isti.cnr.it/~nardini, h-index (Google Scholar), 24) is a researcher with the National Research Council of Italy. He received the Ph.D. in Information Engineering from the University of Pisa in 2011. His research interests focus on Web Information Retrieval (IR), Data Mining (DM), and Machine Learning. He served as a program committee member of several top-level conferences of IR and DM. He authored more than 50 papers in peer-reviewed international journals, conferences and other venues.

Salvatore Trani is a researcher with the Italian National Research Council of Italy. He received the Ph.D. in Computer Science from the University of Pisa in 2017. His research interests focus on Information Retrieval (IR), Machine Learning (ML) and Semantic Enrichment. He served as a program committee member of several top-level conferences of IR and ML. He authored more than 15 papers in peer-reviewed international journals, conferences and other venues.

**Final Exam**: Yes. Three possibilities (students can choose their preferred one): 1) final written exam (open questions), 2) seminar (30 min.) presenting topics related to the course, 3) project focusing on course topics.

**Room and Schedule**

Room: *Aula Riunioni del Dipartimento di Ingegneria dell'Informazione, Via G. Caruso 16, Pisa – Ground Floor*

- 18/06/2024: 9:00 - 13:00
- 19/06/2024: 9:00 - 13:00
- 20/06/2024: 9:00 - 13:00
- 21/06/2024: 9:00 - 13:00