Course guide
270657 - IR - Information Retrieval

Last modified: 12/07/2021

Unit in charge: Barcelona School of Informatics
Teaching unit: 723 - CS - Department of Computer Science.

Degree: MASTER'S DEGREE IN INNOVATION AND RESEARCH IN INFORMATICS (Syllabus 2012). (Optional subject).

Academic year: 2021 **ECTS Credits:** 6.0 **Languages:** English

LECTURER

Coordinating lecturer: RAMON FERRER CANCHO

Others: Primer quadrimestre:
RAMON FERRER CANCHO - 10

PRIOR SKILLS

Those assumed for admision to MIRI plus those provided by the common learning phase.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CEC2. Capacity for mathematical modelling, calculation and experimental design in engineering technology centres and business, particularly in research and innovation in all areas of Computer Science.
CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.

Generical:
CG1. Capability to apply the scientific method to study and analyse of phenomena and systems in any area of Computer Science, and in the conception, design and implementation of innovative and original solutions.
CG3. Capacity for mathematical modeling, calculation and experimental designing in technology and companies engineering centers, particularly in research and innovation in all areas of Computer Science.
CG5. Capability to apply innovative solutions and make progress in the knowledge to exploit the new paradigms of computing, particularly in distributed environments.

Transversal:
CTR4. INFORMATION LITERACY: Capability to manage the acquisition, structuring, analysis and visualization of data and information in the area of informatics engineering, and critically assess the results of this effort.
CTR5. APPROPIATE ATTITUDE TOWARDS WORK: Capability to be motivated by professional achievement and to face new challenges, to have a broad vision of the possibilities of a career in the field of informatics engineering. Capability to be motivated by quality and continuous improvement, and to act strictly on professional development. Capability to adapt to technological or organizational changes. Capacity for working in absence of information and/or with time and/or resources constraints.
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.

Basic:
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB8. Capability to communicate their conclusions, and the knowledge and rationale underpinning these, to both skilled and unskilled public in a clear and unambiguous way.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.

TEACHING METHODOLOGY

Sessions of theory + problemes of 3 sessions per week. The 2 hours of each session are theoretical expositions, and the third one is devoted to joint exercise solving. For each session, the student will have to deliver solutions to a few problems proposed but not solved in the previous session.

Laboratory sessions of 1 hour per week. For many of the sessons, the student will have to deliver a report of the work done and obtained results after about two weeks.

The working of each type of session is described in the "Activities" session.

Furthermore, at the end of the course each student must present to instructors and fellow students a scientific paper related to the course topic, in the format of a conference presentation. Near week 8 of the course, a list of papers will be made public, from which each student can choose one, or alternatively propose a paper of his/her choice, to be approved by the instructors. The date and time range for the presentations will be announced with at least 2 months time, and the schedule within the chosen day at least 1 week time.

LEARNING OBJECTIVES OF THE SUBJECT

STUDY LOAD

Type	Hours	Percentage
Hours large group	54,0	36.00
Self study	96,0	64.00

Total learning time: 150 h

CONTENTS

Introduction

Description:
Need of search and analysis techniques of massive information. Search and analysis vs. databases. Information retrieval process. Preprocessing and lexical analysis.

Models of information retrieval

Description:
Formal definition and basic concepts: abstract models of documents and query languages. Boolean model. Vector model. Latent Semantic Indexing.

Implementation: Indexing and searching

Description:
Inverse and signature files. Index compression. Example: Efficient implementation of the rule of the cosine measure with tf-idf. Example: Lucene.

Evaluation in information retrieval

Description:
Recall and precision. Other performance measures. Reference collections. Relevance feedback and query expansion.

Web search

Description:
Ranking and relevance in the web. The PageRank algorithm. Crawling. Architecture of a simple web search system.

Architecture of massive information processing systems

Description:
Scalability, high performance, and fault tolerance: the case of massive web searchers. Distributed architectures. Example: Hadoop.

Network analysis

Description:
Descriptive parameters and characteristics of networks: degree, diameter, small-world networks, among others. Algorithms on networks: clustering, community detection and detection of influential nodes, reputation, among others.

Information Systems based on massive information analysis. Combination with other technologies.

Description:
Search Engine Optimization. Joint use of IR techniques with Data Mining and Machine Learning. Recommender Systems.

ACTIVITIES

Theoretical development of topics 1 to 8 of the course

Description:
The student will attend the instructor's presentation and actively participate in the initial discussion of the challenge to be solved in that session.

Full-or-part-time: 52h
Theory classes: 26h
Self study: 26h

Exercises on topics 1 to 8 of the course

Description:
In each session, the instructor proposes a number of exercises (say, 4 to 7) on the topic just covered in theory. Next, a few of the problems (say, 3) are solved jointly. Students must solve the rest of the exercises and deliver them by the start of next session. A part of the session is devoted to discussing the possible questions that may have appeared while solving the problems pending from the last session.

Full-or-part-time: 39h
Practical classes: 13h
Self study: 26h

Laboratory work on topics 1 to 8

Description:
The teacher will describe a practical work to be carried out related with the topics most recently covered. This may be a data analysis task, the implementation of an algorithm seen in class, or proposing a solution for an Information Retrieval scenario. The student completes the work as much as possible in class, although occasionally some additional time may be necessary. In many cases the student will have to produce a report on the work done and results obtained, to be delivered within some clearly stated deadline (say, 2 weeks).

Full-or-part-time: 26h
Laboratory classes: 13h
Self study: 13h

Final exam

Description:
Final exam on the contents of the whole course

Full-or-part-time: 18h
Guided activities: 3h
Self study: 15h

Study and presentation of a scientific paper

Description:
Study and presentation of a scientific paper related to the course topic

Full-or-part-time: 13h
Guided activities: 3h
Self study: 10h

GRADING SYSTEM

Define:

- NF as the grade of the final exam
- NE the grade of exercise assignments
- NL the grade of lab reports
- NA the grade from the presentation of a scientific article

(all in the range 0..10).

Then the final course grade is 0.3*NF + 0.25*NL + 0.25*NE + 0.2*NA.

BIBLIOGRAPHY

Basic:

- Manning, C.D.; Raghavan, P.; Schütze, H. Introduction to information retrieval. Cambridge University Press, 2008. ISBN 9780521865715.
- Croft, W.B.; Metzler, D.; Strohman, T. Search engines: information retrieval in practice. Pearson, 2010. ISBN 9780131364899.
- Russell, M.A.; Klassen, M. Mining the social web: data mining Facebook, Twitter, LinkedIn, Instagram, Github, and more. 3rd ed. O'Reilly Media, 2018. ISBN 9781491973509.
- McCandless, M.; Hatcher, E.; Gospodnetic, O. Lucene in action. 2nd ed. Manning, 2010. ISBN 9781933988177.
- Baeza-Yates, R.; Ribeiro-Neto, B. Modern information retrieval: the concepts and technology behind search. 2nd ed. Addison-Wesley / Pearson, 2011. ISBN 9780321416919.

RESOURCES

Hyperlink:

- http://www.cs.upc.edu/~IR-MIRI/