Course guide
340455 - REIN-I7P23 - Information Retrieval

Last modified: 30/06/2023

Unit in charge: Vilanova i la Geltrú School of Engineering
Teaching unit: 723 - CS - Department of Computer Science.

Degree: BACHELOR'S DEGREE IN INFORMATICS ENGINEERING (Syllabus 2018). (Optional subject).

Academic year: 2023 **ECTS Credits:** 6.0 **Languages:** Catalan

LECTURER

Coordinating lecturer: Neus Català i Roig

Others: Neus Català i Roig

PRIOR SKILLS

- To know and use comfortably basic concepts of linear algebra, discrete mathematics, probability and statistics.
- To program comfortably in object-based languages, including inheritance between classes.
- To know the main data structures to access information efficiently and their implementations (lists, hashing, trees, graphs, heaps). To be able to use them to build efficient programs. To be able to analyze the execution time and memory used by an algorithm of average difficulty. To have an idea of the difference in time to access main memory and disk.
- To know the main elements of a relational database and SQL-like access language.

REQUIREMENTS

Have passed ESTA, AMEP and DABD courses or at least being enrolled.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
1. CEC07. Ability to learn and develop techniques of computing learning and design and implement applications and systems which use them, including those dedicated to automatic information and knowledge extraction from large data volums.
4. CEIS6. Ability to design appropriate solutions in one or more application domains using software engineering methods that integrate ethical, social, legal and economic aspects.
3. CEIS4. Ability to identify and analyze problems and design, develop, deploy, test and document software solutions based on an adequate knowledge of theories, models and techniques.
2. CEIS1. Ability to develop, to maintain and avaluate programming services and systems which satisfy all requirements of user having a reliable and efficient behavior, being comprehensible to develop and maintain and observe to current rules, applying theory, principals, methods, practices of pragramming engineering.

Transversal:
5. ENTREPRENEURSHIP AND INNOVATION: Knowing about and understanding how businesses are run and the sciences that govern their activity. Having the ability to understand labor laws and how planning, industrial and marketing strategies, quality and profits relate to each other.

TEACHING METHODOLOGY

The methodological approach consists of:
- 2 hours per week of lecture classes in which the teacher presents subject matter to students (theory lectures and problem-solving sessions),
- 2 hours per week in the computer classroom, in which students will do the work specified in the script with the guidance of the teacher.

LEARNING OBJECTIVES OF THE SUBJECT

The amount of information stored digitally in organizations, or collectively on the web, is today large enough to make searching this information a generally complicated task. The field known as "Information Retrieval" finds methods to organize information in such a way that finding information afterwards can be done simply and efficiently.

This course will cover basic keyword-based techniques to search in textual information. The course will also examine search in the web, where hyperlinks can be used not only to direct the search but to assess the interest value of each page - as is the case with the well-known PageRank algorithm. Extensions of these techniques to the cases of personalized search and recommender systems is discussed. Finally, a brief introduction to semantic search and neural models of information retrieval is included.

STUDY LOAD

Type	Hours	Percentage
Hours small group	30,0	20.00
Self study	90,0	60.00
Hours large group	30,0	20.00

Total learning time: 150 h

CONTENTS

1. Introduction

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

Related activities:
Activity 1: Mid-term exam
Activity 3: Laboratory sessions

Full-or-part-time: 11h
Theory classes: 1h 30m
Laboratory classes: 2h 30m
Self study : 7h

2. Models of information retrieval

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

Related activities:
Activity 1: Mid-term exam
Activity 3: Laboratory sessions

Full-or-part-time: 12h
Theory classes: 1h 30m
Laboratory classes: 3h 30m
Self study : 7h

3. 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: ElasticSearch.

Related activities:
Activity 1: Mid-term exam
Activity 3: Laboratory sessions

Full-or-part-time: 10h
Theory classes: 0h 30m
Laboratory classes: 2h 30m
Self study : 7h

4. Evaluation in information retrieval

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

Related activities:
Activity 1: Mid-term exam
Activity 3: Laboratory sessions

Full-or-part-time: 10h
Theory classes: 0h 30m
Laboratory classes: 2h 30m
Self study : 7h

5. Web search

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

Related activities:
Activity 2: Second partial exam
Activity 3: Laboratory sessions

Full-or-part-time: 16h
Theory classes: 3h
Laboratory classes: 6h
Self study : 7h

6. Architecture of massive information processing systems

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

Related activities:
Activity 2: Second partial exam
Activity 3: Laboratory sessions

Full-or-part-time: 12h 30m
Theory classes: 3h
Laboratory classes: 6h
Self study : 3h 30m

7. Information systems based on massive information analysis

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

Related activities:
Activity 2: Second partial exam
Activity 3: Laboratory sessions

Full-or-part-time: 10h 30m
Theory classes: 2h
Laboratory classes: 5h
Self study : 3h 30m

8. Semantic search and neural models of information retrieval

Description:
Semantic search: how to make searches include meanings and contexts, not just keywords. Word embeddings and sentence embeddings. Neural models of information retrieval. Re-ranking.

Related activities:
Activity 2: Second partial exam
Activity 3: Laboratory sessions

Full-or-part-time: 15h
Theory classes: 2h
Laboratory classes: 6h
Self study : 7h

GRADING SYSTEM

The course will include the following evaluation events:
- Reports of laboratory sessions (L).
- A mid-term exam, covering material seen until the exam is done (C1).
- A second partial exam (C2), covering what was not covered in the mid-term exam.

The final grade is computed by the following formula:

0.4*L + 0.3*C1 + 0.3*C2

EXAMINATION RULES.

Reports of laboratory sessions will be delivered online within a time limit for each session.

Mid-term exam and second partial exam are in-person.

BIBLIOGRAPHY

Basic:
- Russell, Matthew A; Klassen, Mikhail. Mining the social web : data mining Facebook, Twitter, LinkedIn, Instagram, Github, and more [on line]. 3rd ed. Sebastopol, [California]: O'Reilly Media, 2018 [Consultation: 14/02/2024]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=5611114. ISBN 9781491973509.
- Baeza-Yates, Ricardo ; Ribeiro-Neto, Berthier. Modern information retrieval : the concepts and technology behind search. 2nd ed. Harlow [etc.]: Addison-Wesley, 2011. ISBN 9780321416919.
- Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich. Introduction to information retrieval [on line]. New York: Cambridge University Press, 2008 [Consultation: 25/03/2022]. Available on: https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html. ISBN 9780521865715.
- Croft, W. Bruce; Metzler, Donald; Strohman, Trevor. Search engines : information retrieval in practice. Boston [etc.]: Pearson, 2010. ISBN 9780131364899.

RESOURCES

Other resources:
Web links:

- An Introduction to Neural Information Retrieval, by Bhaskar Mitra and Nick Craswell (https://arxiv.org/abs/1705.01509)

- The Anatomy of a Large-Scale Hypertextual Web Search Engine, by Sergey Brin and Lawrence Page (http://infolab.stanford.edu/backrub/google)

- The Heart of the Elastic Stack (https://www.elastic.co/products/elasticsearch)