Last update: 12-07-2019

340455 - REIN-I7P23 - Information Retrieval

Coordinating unit: 340 - EPSEVG - Vilanova i la Geltrú School of Engineering

Teaching unit: 723 - CS - Department of Computer Science

Academic year: 2019

Degree: BACHELOR'S DEGREE IN INFORMATICS ENGINEERING (Syllabus 2018). (Teaching unit Optional)
BACHELOR'S DEGREE IN INFORMATICS ENGINEERING (Syllabus 2010). (Teaching unit Optional)

ECTS credits: 6 Teaching languages: Catalan

Teaching staff

Coordinator: Neus Català Roig

Others: Neus Català Roig

Opening hours

Timetable: See the current office hours in the EPSEVG people list:

https://web3.epsevg.upc.edu/coneix-lepsevg/directori-epsevg

Prior skills

- To know and use comfortably basic concepts of linear algebra, discrete mathematics, probability and statistics.

- To program comfortably in object-oriented 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.

340455 - REIN-I7P23 - Information Retrieval

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 case of Social Networks where interactions among users can provide very useful information will be seen. Finally, the course will study ways in which these techniques can be exploited for the benefit of specific organizations.

Study load

Total learning time: 150h	Hours large group:	30h	20.00%
	Hours medium group:	0h	0.00%
	Hours small group:	30h	20.00%
	Guided activities:	0h	0.00%
	Self study:	90h	60.00%

Content

1. Introduction

Learning time: 11h

Theory classes: 1h 30m
Laboratory classes: 2h 30m
Self study : 7h

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: Final exam
Activity 4: Laboratory sessions

2. Models of information retrieval

Learning time: 12h

Theory classes: 1h 30m
Laboratory classes: 3h 30m
Self study : 7h

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: Final exam
Activity 4: Laboratory sessions

3. Implementation: Indexing and searching

Learning time: 10h

Theory classes: 0h 30m
Laboratory classes: 2h 30m
Self study : 7h

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: Final exam
Activity 4: Laboratory sessions

4. Evaluation in information retrieval

Learning time: 10h

Theory classes: 0h 30m
Laboratory classes: 2h 30m
Self study : 7h

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

Related activities:
Activity 1: Mid-term exam
Activity 3: Final exam
Activity 4: Laboratory sessions

5. Web search

Learning time: 16h

Theory classes: 3h
Laboratory classes: 6h
Self study : 7h

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: Final exam
Activity 4: Laboratory sessions

6. Architecture of massive information processing systems

Learning time: 12h 30m

Theory classes: 3h
Laboratory classes: 6h
Self study : 3h 30m

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: Final exam
Activity 4: Laboratory sessions

7. Network analysis	Learning time: 15h Theory classes: 2h Laboratory classes: 6h Self study : 7h

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.

Related activities:
Activity 2: Second partial exam
Activity 3: Final exam
Activity 4: Laboratory sessions

8. Information systems based on massive information analysis.	Learning time: 10h 30m Theory classes: 2h Laboratory classes: 5h Self study : 3h 30m

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: Final exam
Activity 4: Laboratory sessions

Qualification system

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

Re-evaluation: There will be a Final Exam (F) covering the whole course. The mark of the Final Exam will substitute the previous grade obtained from re-evaluable tasks (C1 and C2) only if it is greater than the latter.

The final grade is computed by the following formula:

0.4*L + 0.3*C1 + 0.3*C2

In case of re-evaluation, the final grade is computed by the following formula:

0.4*L + 0.6*F

Regulations for carrying out activities

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

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

Bibliography

Basic:

Baeza-Yates, Ricardo ; Ribeiro-Neto, Berthier. Modern information retrieval : the concepts and technology behind search. 2nd. Harlow [etc.]: Addison-Wesley, 2011. ISBN 978-0321416919.

Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich. Introduction to information retrieval [Recurs electrònic] [on line]. Cambridge [etc.]: Cambridge University Press, 2008 [Consultation: 27/11/2014]. Available on: <http://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.

Russell, Matthew A.. Mining the social web: data mining Facebook, Twitter, LinkedIn, Google+, Github, and more [on line]. 2nd ed. Sebastopol, [California]: O'Reilly, 2013 [Consultation: 04/03/2015]. Available on: <http://site.ebrary.com/lib/upcatalunya/docDetail.action?docID=10779158>. ISBN 9781449367619.

Others resources:

Web links:
- Three and a half degrees of separation, by Smriti Bhagat, Moira Burke, Carlos Diuk, Ismail Onur Filiz, Sergey Edunov (https://research.fb.com/three-and-a-half-degrees-of-separation/?refid)
- 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)