

Course guide

270220 - CAI - Information Retrieval and Analysis

Last modified: 19/07/2023

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

Degree: BACHELOR'S DEGREE IN DATA SCIENCE AND ENGINEERING (Syllabus 2017). (Compulsory subject).

Academic year: 2023 **ECTS Credits:** 6.0 **Languages:** Catalan, Spanish, English

LECTURER

Coordinating lecturer: MARTA ARIAS VICENTE

Others: Primer quadrimestre:
MARTA ARIAS VICENTE - 11, 12, 13

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:

CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.

CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

CE6. Build or use systems of processing and comprehension of written language, integrating it into other systems driven by the data. Design systems for searching textual or hypertextual information and analysis of social networks.

CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.

Generical:

CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

CG3. Work in multidisciplinary teams and projects related to the processing and exploitation of complex data, interacting fluently with engineers and professionals from other disciplines.

CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

CG5. To be able to draw on fundamental knowledge and sound work methodologies acquired during the studies to adapt to the new technological scenarios of the future.

Transversal:

CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.

CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

Basic:

CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.

CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

TEACHING METHODOLOGY

Classes "de teoria" expositives per part del professor. Es proposaran un cert nombre d'exercicis a resoldre fora de classe per a la propera sessió.

Classes "de teoria" dedicades a la resolució. Es comentaran en comú les solucions dels exercicis proposats a la/les sessions precedents. S'esperarà que els estudiants hagin intentat resoldre'ls.

Classes "de laboratori": A partir d'un guió que rebran al principi de la sessió, els estudiants duran a terme alguna tasca amb ordinador per consolidar els conceptes vistos a les classes de "teoria". Típicament serà l'implementació i experimentació amb algun algorisme, o l'anàlisi d'algun conjunt de dades.

LEARNING OBJECTIVES OF THE SUBJECT

1. Describe different models for evaluating similarity between texts, and how they apply to textual search. Decide which of the models is best suited to a specific scenario involving text search. Implement the models from scratch (in a very basic system) or on a highly scalable text indexing system.

2. Describe the advantages, in order to carry out effective searches, of using the information given by links in hyperlink structures, such as the web, digital social networks, and the semantic web. Describe the main parameters used to characterize these linked structures. Reproduce the most commonly used algorithms to establish importance in these structures (e.g. pagerank), to discover structure in them (e.g. community discovery) and to improve search results proposed by a user. Implement these algorithms from scratch in a very basic system, or on top of massive data processing systems so that they can scale.

Translated with www.DeepL.com/Translator

3. Evaluate the effectiveness of search systems in complex systems, describing it in terms of hard measures such as "recall" and "accuracy" but also in terms of soft measures such as user satisfaction, novelty and task completion. Adapt the operation and presentation of information search systems with feedback from the user experience methodically collected.

4. Define the problem of the recommendation and the differences with other problems related to information previously stored (search, learning, ...). Describe the main approaches to the problem of item recommendations and the advantages and disadvantages of each one. Describe the main algorithms of each of the approaches. Be able to implement basic versions from scratch, or advanced versions on top of massive data processing systems. Evaluate the effectiveness of recommendation systems, both in terms of hard measures and soft measures such as user satisfaction. Decide on the most appropriate forms of recommendation to simple real scenarios, including the characterization of potential users.

Translated with www.DeepL.com/Translator

5. Use known algorithmic paradigms to deal with data problems characterized by high volume and high speed. They include: streaming algorithms that treat data flows with little time per element, and little memory. Algorithms to answer proximity questions, particularly with geolocalized information. Algorithms that use sampling to draw reliable conclusions about large volumes of data. Integration of the techniques seen in the rest of the course with algorithmic techniques of other subjects, such as "machine learning", "clustering" and "pattern mining". Techniques for dealing with sensitive data, such as anonymization and privacy-preserving machine learning. "Consistent and distributed caching."

Translated with www.DeepL.com/Translator

6. Integrate the techniques described in the previous objectives into a small but realistic project. Have the ability to design the architecture of a complex system and choose the techniques and technologies previously seen during the course to be applied. The objective is not to finalize the implementation of the system, but to arrive at a level of design detail that would allow a programming team to commission its completion.

7. To evaluate in a basic way the implications of the systems that are learned to build in the subject in terms of privacy, security, ethics and people's rights. It is understood by "in an elementary way" to be able to detect that these implications are significant enough to seek the opinion of an expert in the matter, particularly in relation to the RGPD and the need to carry out risk and impact analysis.



STUDY LOAD

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

Total learning time: 150 h

CONTENTS

Search and analysis of text information

Description:

Models booleà i vectorial. Cerca basada en paraules clau. Preprocés dels textos. Indexació. Avaluació d'estratègies de cerca. Formació de grups i classificació de textos. Models generatius (LSI, LDA).

Search and analysis in linked structures

Description:

La web: Algorismes d'avaluació en estructures hiperenllaçades. "Crawling" i "scraping". Xarxes socials: Mesures de centralitat. Comunitats. Influència. Web semàntica.

Recommendation

Description:

Sistemes recomanadors. Recomanació basada en contingut i recomanació basada en la comunitat ("collaborative filtering"). Consideracions pràctiques.

Massive data algorithms

Description:

Resums (sketches) i fluxos de dades (streaming). Mostratge (sampling). Preguntes de proximitat. Dades geolocalitzades. "Caching" consistent i distribuït. Tractament de dades sensibles: anonimització, "end-to-end encryption" i "privacy-preserving machine learning"

ACTIVITIES

Activitat sobre el contingut "Cerca i anàlisi d'informació textual"

Specific objectives:

1, 3, 6, 7

Related competencies :

CG3. Work in multidisciplinary teams and projects related to the processing and exploitation of complex data, interacting fluently with engineers and professionals from other disciplines.

CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

CG5. To be able to draw on fundamental knowledge and sound work methodologies acquired during the studies to adapt to the new technological scenarios of the future.

CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.

CE6. Build or use systems of processing and comprehension of written language, integrating it into other systems driven by the data. Design systems for searching textual or hypertextual information and analysis of social networks.

CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.

CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.

CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

CB3. That students have the ability to gather and interpret relevant data (usually within their area of study) to make judgments that include a reflection on relevant social, scientific or ethical issues.

CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of study.

Full-or-part-time: 24h

Theory classes: 6h

Laboratory classes: 6h

Self study: 12h

Activitat sobre el contingut "Cerca i anàlisi en estructures enllaçades"

Specific objectives:

2, 6, 7

Related competencies :

CG3. Work in multidisciplinary teams and projects related to the processing and exploitation of complex data, interacting fluently with engineers and professionals from other disciplines.

CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

CG5. To be able to draw on fundamental knowledge and sound work methodologies acquired during the studies to adapt to the new technological scenarios of the future.

CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.

CE6. Build or use systems of processing and comprehension of written language, integrating it into other systems driven by the data. Design systems for searching textual or hypertextual information and analysis of social networks.

CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.

CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.

CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.

CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

Full-or-part-time: 24h

Theory classes: 6h

Laboratory classes: 6h

Self study: 12h

Activitat sobre el tema "Recomanació"

Specific objectives:

4, 6, 7

Related competencies :

CG3. Work in multidisciplinary teams and projects related to the processing and exploitation of complex data, interacting fluently with engineers and professionals from other disciplines.

CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

CG5. To be able to draw on fundamental knowledge and sound work methodologies acquired during the studies to adapt to the new technological scenarios of the future.

CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.

CE6. Build or use systems of processing and comprehension of written language, integrating it into other systems driven by the data. Design systems for searching textual or hypertextual information and analysis of social networks.

CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.

CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.

CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.

CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

Full-or-part-time: 16h

Theory classes: 4h

Laboratory classes: 4h

Self study: 8h

Activitat sobre el contingut "Algorismes per a dades massives"

Specific objectives:

5, 6, 7

Related competencies :

CG3. Work in multidisciplinary teams and projects related to the processing and exploitation of complex data, interacting fluently with engineers and professionals from other disciplines.

CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

CG5. To be able to draw on fundamental knowledge and sound work methodologies acquired during the studies to adapt to the new technological scenarios of the future.

CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.

CE6. Build or use systems of processing and comprehension of written language, integrating it into other systems driven by the data. Design systems for searching textual or hypertextual information and analysis of social networks.

CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.

CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.

CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.

CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

Full-or-part-time: 34h

Theory classes: 8h

Laboratory classes: 8h

Self study: 18h

Integració. Construcció de sistemes reals. Implicacions en privacitat, seguretat i drets de les persones.

Specific objectives:

6, 7

Related competencies :

CG3. Work in multidisciplinary teams and projects related to the processing and exploitation of complex data, interacting fluently with engineers and professionals from other disciplines.

CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

CG5. To be able to draw on fundamental knowledge and sound work methodologies acquired during the studies to adapt to the new technological scenarios of the future.

CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.

CE6. Build or use systems of processing and comprehension of written language, integrating it into other systems driven by the data. Design systems for searching textual or hypertextual information and analysis of social networks.

CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.

CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.

CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.

CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

Full-or-part-time: 16h

Theory classes: 4h

Laboratory classes: 4h

Self study: 8h

Partial exam

Specific objectives:

1, 2, 3

Related competencies :

CG3. Work in multidisciplinary teams and projects related to the processing and exploitation of complex data, interacting fluently with engineers and professionals from other disciplines.

CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

CG5. To be able to draw on fundamental knowledge and sound work methodologies acquired during the studies to adapt to the new technological scenarios of the future.

CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.

CE6. Build or use systems of processing and comprehension of written language, integrating it into other systems driven by the data. Design systems for searching textual or hypertextual information and analysis of social networks.

CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.

CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.

CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.

CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

Full-or-part-time: 13h

Guided activities: 3h

Self study: 10h

Final exam

Specific objectives:

1, 2, 3, 4, 5, 7

Related competencies :

CG3. Work in multidisciplinary teams and projects related to the processing and exploitation of complex data, interacting fluently with engineers and professionals from other disciplines.

CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

CG5. To be able to draw on fundamental knowledge and sound work methodologies acquired during the studies to adapt to the new technological scenarios of the future.

CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.

CE6. Build or use systems of processing and comprehension of written language, integrating it into other systems driven by the data. Design systems for searching textual or hypertextual information and analysis of social networks.

CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.

CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.

CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.

CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

Full-or-part-time: 15h

Guided activities: 3h

Self study: 12h

GRADING SYSTEM

P = partial take-home exam mark, mid term.

F = final exam mark.

L = lab session reports mark.

Final will be computed as 20% P + 40% F + 40% L.

The grade assigned to the "competencia transversal" CT6 (autonomous learning) will be computed from exam responses and/or information reflected on project reports from a topic proposed by the instructor that students will have to learn on their own.



BIBLIOGRAPHY

Basic:

- Leskovec, J.; Rajaraman, A.; Ullman, J.D. Mining of massive datasets. 3rd ed. Cambridge: Cambridge University Press, 2020. ISBN 9781108476348.
- Stephens-Davidowitz, S. Everybody lies : what the internet can tell us about who we really are. Bloomsbury Publishing, 2018. ISBN 9781408894736.
- Baeza-Yates, R.; Ribeiro-Neto, B. Modern information retrieval: the concepts and technology behind search. 2nd ed. Addison-Wesley / Pearson, 2011. ISBN 9780321416919.