Universitat Autònoma de Barcelona

2020/2021

Data Retreival and Storage

Code: 104851
ECTS Credits: 6

Degree	Type	Year	Semester
2503852 Applied Statistics	FB	1	2

The proposed teaching and assessment methodology that appear in the guide may be subject to changes as a result of the restrictions to face-to-face class attendance imposed by the health authorities.

Contact

Name: Marc Vallribera Ros
Email: Marc.Vallribera@uab.cat

Use of Languages

Principal working language: catalan (cat)
Some groups entirely in English: No
Some groups entirely in Catalan: Yes
Some groups entirely in Spanish: No

Prerequisites

Knowledge about logical operations.
Knowledge about sets and relationships between sets.
Basic knowledge of Python.

Objectives and Contextualisation

In this subject, the basic concepts of Databases (DB) necessary for both DB designer and user level are introduced, as well as the mechanisms for obtaining information from the Internet through Web Scraping and store it in a DB.

Competences

- Correctly use a wide range of statistical software and programming languages, choosing the best one for each analysis, and adapting it to new necessities.
- Select the sources and techniques for acquiring and managing data for statistical processing purposes.
- Students must be capable of applying their knowledge to their work or vocation in a professional way and they should have building arguments and problem resolution skills within their area of study.
- Students must have and understand knowledge of an area of study built on the basis of general secondary education, and while it relies on some advanced textbooks it also includes some aspects coming from the forefront of its field of study.
- Use quality criteria to critically assess the work done.

Learning Outcomes

1. Correctly identify the types of data and measurements.
2. Critically assess the work done on the basis of quality criteria.
3. Identify the advantages and disadvantages of the internet as a major source of information in statistics.
4. Manage a database.
5. Perform basic information-purging operations.

6. Students must be capable of applying their knowledge to their work or vocation in a professional way and they should have building arguments and problem resolution skills within their area of study.
7. Students must have and understand knowledge of an area of study built on the basis of general secondary education, and while it relies on some advanced textbooks it also includes some aspects coming from the forefront of its field of study.
8. Understand the computer algorithms used to manage a database and the SQL language.
9. Use suitable information sources for each type of applied study.

Content

1. Introduction to databases
    1. Information systems and databases
    2. Databases
        1. Concepts
        2. Features
        3. Database evolution
    3. Definition and characteristics of an DBMS
    4. Architecture of the DBMS
    5. Main DBMS
    6. BD application development phases
    7. BD Design Stages
        1. Conceptual design
        2. Logical design
        3. Application design
2. The Entity-Relationship Model (E-R)
    1. The E-R diagram
    2. Entities, attributes, interrelationships
    3. Interrelationship attributes
    4. Dependence on existence and participation
3. The relational model
    1. Relationship or table concept, attributes, tuples, domains, primary and external keys
    2. Domain restrictions, key integrity, and reference
    3. Transforming the E/R model to relational
    4. Relational algebra operators
4. Database implementation
    1. Structured Query Language (SQL)
    2. Data processing
    3. Data query
    4. Data Base management with SQL
    5. Working with SQLite databases
5. HTML and Regular Expressions basics
    1. Web page code structure
    2. HTML and CSS tags and attributes
    3. Text search using Regular Expressions
    4. Special characters, sets, groups and repetitions
6. Collecting and storing data from web pages
    1. Introduction to WebScrapping Tools
    2. Programming Web Scraping tools using Python
    3. Searching and obtaining information with Regular Expressions
    4. Searching and obtaining information with Beautiful Soup
    5. Database storage
    6. Exporting results in comma separated values files

Methodology

Theory
Classes are taught through master classes with transparencies. These transparencies are accessible, and the students can obtain them from the Virtual Campus.

Continuous assessment
There will be 3 deliveries (individual) so that the student can prove that is acquiring the knowledge that is explained in the class. The delivery will be done through the Moodle on the Virtual Campus.

Proposed problems
during the course, a list of problems will be provided, about the most practical topics of the subject, so that the student can acquire and/or consolidate their knowledge of the different stages in the design, implementation and exploitation of the databases.

Preparation of the practices
The student must have read and prepared the practices in order to be able to do them within the established schedule of practices and at home.

Practices
The objective of the lab sessions is to give a broad vision of the databases, from management and creation to the connection with an application that allows you to consult and modify the data. Students will have to acquire competences in the creation, management and manipulation of databases, as well as obtaining information from the Internet, and the storage of that data in the database. Throughout these lab sessions, the teacher will supervise and guide every group of students during the process.

Activities

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Theory lessons	26	1.04	2, 8, 5, 4, 1, 3, 6, 9
Type: Supervised			
Continuous Evaluation Deliveries	9	0.36	2, 8, 4, 1, 9
Practices	36	1.44	2, 8, 5, 4, 1, 6, 9
Type: Autonomous			
Books reading	20	0.8	8, 1, 3, 9
Practices preparation	10	0.4	2, 8, 5, 4, 1, 6, 9
Proposed problems	23	0.92	1, 3, 6, 9
Study	15	0.6	2, 8, 5, 4, 1, 3, 6, 9

Assessment

70% of the course grade will be based on practices mark and a final exam, which can be recovered. The remaining 30% will be assessed through continuous evaluation deliveries. All notes listed below are on 10.

The final grade will be: Final mark = 0.4 * Exam mark + 0.3 Practices mark + 0.3 * Continuous marks

Passing the course requires passing the practices and exam separately.

Exam (40%): The main exam of the course will be held on the last day of class. The recovery exam will be held on the day reserved for this subject within the exams schedule.

Practices (30%): There will be a evaluated delivery of the sessions.

At the end of the course, the practices can be re-evaluated, with a special delivery. The maximum grade that can be obtained in the re-evaluation of practices will be 5.

Continuous evaluation (30%): The continuous evaluationmark will be obtained from the problems that will be delivered during the course. The specific form and the days that the deliveries of the problems will be notified with prior notice on the Virtual Campus of the subject. The continuous evaluation mark is not recoverable.

Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Excercise delivery (Continuous Evaluation)	30%	3	0.12	2, 8, 5, 4, 1, 7, 6
Final exam	40%	3	0.12	2, 8, 5, 4, 1, 3, 7, 6
Practices delivery	30%	5	0.2	2, 8, 5, 4, 1, 3, 7, 6, 9

Bibliography

A. Silberschatz, H.F. Korth, S. Sudarshan (2006), *Fundamentos de Bases de Datos*, McGraw-Hill
Ryan Mitchell (2018), *Web Scraping with Python*, O'Reilly Media, Inc, USA [ISBN 1491985577]
Michael Heydt (2018) *Python Web Scraping Cookbook*, Packt Publishing [ISBN 1787285219]
Ian Mackie (2020) A Begginners Guide to Python 3 Programming