Data Harvesting

Academic Year: (2022 / 2023) **Review date: 18/05/2022 20:20:07**

Department assigned to the subject: Computer Science and Engineering Department

Coordinating teacher: GENOVA FUSTER, GONZALO

Type: Compulsory ECTS Credits : 3.0

Year : 1 Semester : 2

REQUIREMENTS (SUBJECTS THAT ARE ASSUMED TO BE KNOWN)

Data Programming (19138)

OBJECTIVES

Core Competences:

- Having and understanding the knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context
- Students know how to apply their acquired knowledge and problem-solving skills in new or unfamiliar settings within broader (or multidisciplinary) contexts related to their field of study.
- Students are able to integrate knowledge and to face the complexity of making judgments based on information that, being incomplete or limited, includes reflections on the social and ethical responsibilities linked to the application of their knowledge and judgments.
- Students know how to communicate their conclusions and the knowledge and ultimate reasons behind them to specialized and non-specialised audiences in a clear and unambiguous way.
- Students have the learning skills that will enable them to continue studying in a way that will be largely self-directed or autonomous.

General Competences:

- Ability to identify, define and formulate social science problems and solve them using computational techniques. This includes the ability to assess all the factors involved, not only technical but also legal.
- Ability to compile and analyze existing knowledge in the different areas of computational social sciences, and to propose possible solutions to the problems raised.
- Ability to address issues raised under new or unfamiliar environments, within the context of computational social sciences.

Specific Competences:

- Ability to use computational tools specific to the computational social sciences at an advanced level.
- Ability to obtain, prepare, integrate and use information from secondary and web sources.

Learning Outcomes:

- Knowledge of the main tools in computational social sciences.
- Knowledge of the general principles of API design and operation, as well as the most common information exchange formats.
- Ability to identify and access online APIs to download social observational data.
- Ability to compile structured databases from unstructured sources.

DESCRIPTION OF CONTENTS: PROGRAMME

1. An introduction to Web Scraping
- What is Web Scraping?
- Types of Web Scraping
- Data formats: XML and HTML
- Practical access to XML and HTML
- Automation for Web Scraping programs
- Selenium and JavaScript based scraping
- Ethical issues with Web Scraping
- Practical exercises

2. Data APIs
- What is an API
- Fundamentals of API communication
- An introduction to the JSON format
- Create your own API (and share it)
- REST architecture
- APIs as a way to share and obtain data (any kind)
- Automation of API requests
- Talking with Databases
- Authentication and ethical access to APIs
- Practical exercises

3. Automation of Data Acquisition
- Why do we need automation?
- Accessing servers
- Technologies for automating programs
- Automating cron jobs
- Logging tasks
- Practical exercises

LEARNING ACTIVITIES AND METHODOLOGY

Training Activities:

- Theoretical-practical classes
- Tutorials
- Group work
- Individual student work
- Partial and final examinations

Teaching Methods:

- Presentations in the professor's lecture room with computer and audiovisual support, in which the main concepts of the subject are developed and a bibliography is provided to complement the students' learning.
- Critical reading of texts recommended by the subject professor: Press articles, reports, manuals and/or academic articles, either for later discussion in class, or to expand and consolidate knowledge of the subject.
- Resolution of practical cases, problems, etc. raised by the professor, either individually or in a group.
- Presentation and discussion in class, under the moderation of the professor, of topics related to the content of the subject, as well as practical case studies.
- Developing pieces of work and reports, individually or in group.

ASSESSMENT SYSTEM

% end-of-term-examination/test:	20
% of continuous assessment (assigments, laboratory, practicals…):	80

- Participation in the class (10%)
- Individual or group work done during the course (70%)
- Final exam (20%)

BASIC BIBLIOGRAPHY

- Barberá, P. & Steinert-Threlkeld, Z. How to use social media data for political science research. In The SAGE handbook of research methods in political science and international relations (Vol. 2, pp. 404-423). , SAGE Publications Ltd, https://dx.doi.org/10.4135/9781526486387, 2020

- Freelon, D. Computational research in the post-API age. , Political Communication, 35(4), 665-668., 2018

- Nyhuis, D. Web data collection: potentials and challenges. In: The SAGE handbook of research methods in political science and international relations (Vol. 2, pp. 387-403). , SAGE Publications Ltd, https://dx.doi.org/10.4135/9781526486387, 2020

- Perriam, J., Birkbak, A., & Freeman, A. Digital methods in a post-API environment. , International Journal of Social Research Methodology, 23(3), 277-290., 2020

ADDITIONAL BIBLIOGRAPHY

- Aydin, O. R Web Scraping Quick Start Guide: Techniques and tools to crawl and scrape data from websites., -, 2018

- Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. Automated data collection with R: A practical guide to web scraping and text mining. , John Wiley & Sons., 2014

BASIC ELECTRONIC RESOURCES

- . Application Programming Interfaces in R: https://sicss.io/2020/materials/day2-digital-trace-data/apis/rmarkdown/Application_Programming_interfaces.html

- . Using APIs to get data: https://cfss.uchicago.edu/notes/application-program-interface/

- . Screen scraping with R: https://cbail.github.io/ids704/screenscraping/rmarkdown/Screenscraping_in_R.html