

Design Computation

#3

Web-Scraping and Data Mining techniques on Geographical Information

Dates: 2 - 4 July

Location: Universidade de Lisboa, Portugal

Organizers: Stefano Fiorito (PhD candidate), Universidade de Lisboa

João Ventura Lopes (PhD candidate), ISCTE - Instituto Universitário de Lisboa

José Beirão (Assist. Prof., PhD), Universidade de Lisboa

Duration: 2,5 days

number of contact hours: 21 ECTS credits: 1,5

Maximum number of participants: 20

Minimum number of participants: 9

Summary:

The Web, is a powerful resource of Knowledge and Information where the data is *Big*, fluid, structured by its nature, and where self learning algorithms interact with the user and between themselves. In this workshop we want to introduce participants to tools and methods that make it possible to automatically collect and learn from this data.

We will teach how to collect data from the World Wide Web and organize information, depicting patterns and extract from these meaningful insights. The exercise will be done on geospatial information and a case study from Portuguese context will be introduced.

Design Computation

Objectives:

- To get a basic understanding of data scraping and data mining procedures for research and academic purposes
- To automatise process of data collection, data preprocessing, visualization, modeling and deploying using self-made scripts and public APIs
- To introduce the students in the use of visual platforms for coding such Jupyter-Notebook with Python as programming language
- To introduce the students to the ecosystem of the Python Data Science libraries
- Exemplify how the processes may be integrated with other tools and workflows

Program:

July 2 (whole day):

- General introduction on Data, Data Types and Data Structures
- General introduction on Data Scraping and the Ethics of scraping
- Introduction to the tools used for the workshop
- Basic refresh on Python (with some snippets of code)
- Definition of the case study, and sharpening of our tool to extract the data
- Storing the data collected and have the first insights in an Exploratory Data Analysis (EDA) approach

July 3 (whole day):

- Create a complete pipeline of code capable to catch the data and store it
- Use Public APIs to add values to our data. (geocoding)
- Structure and shaping the data collected for the Data Mining procedures
- General introduction on Data Mining and Machine Learning
- Introduction to the Python Data Mining libraries used for the workshop
- Exemplifying and exploring the tools and workflows with standard datasets

July 4 (morning):

- Modeling the collected data for knowledge discovery and prediction with supervised and unsupervised learning techniques
- Deploying, mapping and visualizing data patterns and the insights gained in the process

Design Computation

Prerequisites:

There are no prerequisites required; we only ask from our participants to be intrigued in learning more about these subjects.

Even though this class will be open to everybody, during the workshop we will use Python as program language for scripting. The platform used for the entire workshop will be Anaconda with Jupyter-Notebook, Qgis, and eventually a text editor and Rhinoceros3D/Grasshopper algorithm design tool.

Requirements:

Projector, a computer room in the CIFA lab, with preloaded softwares Anaconda, Qgis and Rhinoceros3D with Grasshopper plugin installed.