3rd term 2017-2018

Masterclass on Web Scraping and Text Mining

From 14 to 18 May 2018 at 9:00-13:00 in Seminar room 3 (Badia Fiesolana)

Mentoring Professor: Elias Dinas (EUI)

Instructor: Paulo Serôdio (Oxford University)

Please register online.

Contact: Jennifer.Dari@eui.eu

Outline

This course will introduce students to the data science fundamentals of extraction, processing and classification of web content. It will review current methods for automated web scraping, natural language processing for parsing unstructured data and machine learning algorithms for textual data. With this in mind, the first part of the course will provide an in-depth survey of different structures and features of web content (XML, JSON, HTML, CSS-tags and XPATH) and cover the main tools for harvesting, extracting and processing the data retrieved into structured formats, using static and dynamic web pages and APIs. In a second stage, we will explore applications of machine learning algorithms to the parsed data, with a particular focus on text analysis. Under the umbrella of supervised and unsupervised learning, the course will cover traditional approaches to content analysis and dictionary-based methods, machine learning algorithms for classification, scaling methods and topic modeling. Our goal is to help students automate the extraction of online content, parse the unstructured data into formats amenable to analysis and produce quantities of interest using classification and data reduction methods, using text as data for the most part. The course will be taught in **R**, but we may also touch upon **Python** libraries for particular applications.

Schedule

Monday	**14 May**	9:00-13:00	(Seminar Room 3, Badia Fiesolana)
Tuesday	**15 May**	9:00-13:00	(Seminar Room 3, Badia Fiesolana)
Wednesday	**16 May**	9:00-13:00	(Seminar Room 3, Badia Fiesolana)
Thursday	**17 May**	9:00-13:00	(Seminar Room 3, Badia Fiesolana)
Friday	**18 May**	9:00-13:00	(Seminar Room 3, Badia Fiesolana)