**Web scrapping: Getting data from the web with R**

**Taught by**: Alex Sánchez. Genetics Microbiology and Statistics Department. Faculty of Biology. University of Barcelona.

**Language**: (English/Spanish/Catalan)

**Description**

An important aspect when dealing with data is that often they are found in the web in formats that require some preprocessing before being analyzed. This course will explore techniques to understand these formats so that you can retrieve data from the web and extract the desired information. The first part introduces the most   
common web technologies, their relationship and some tools to manipulate   
and extract the information. Some most common formats for storing web information (HTML, XML, JSON) are presented, as well as tools to extract it, as XPath and CSS selectors. Finally we introduce some R package suitable to process Web information.

**Objectives**

Specifically at the end of the course students should:

* Be familiar with the main technologies to deal with information stored in   
  the web.
* Be able to recognize the different formats that can be used for storage.
* Know how to extract information from these formats using specific R packages.

**Contents**

1. Introducing Web technologies. Web scrapping and web scrapping projects.
2. Data representation in the web HTML, XML, JSON. Other technologies.
3. Regular expressions for data manipulation.
4. Parsing HTML and XML. Using CSS selectors and XPath.
5. Case study: Scraping Twitter for Sentiment Analysis.

**Evaluation**

It will be based on a practical exercise. Each student will suggest and implement a scrapping task, of a minimal complexity, consistent with the course contents. Some proposals will be available for those who prefer a guided exercise.

Altres qüestions

Pressupost

En principi cap apart de la remuneració.   
Si hi ha la possibilitat de comprar un parell de llibres es podrien afegir al pressupost.

Ordre de preferència per part del profesor

* Preferentment torn de tarda
* Preferentment semana 2 o 1. La tercera semana puc estar en un comité avaluador a Madrid (està pendent de confirmar)