Syllabus

Coding 2: Web Scraping with R, from zero to hero

- **Instructor:** Mihály Orsós (orsosm@ceu.edu – office hours by appointment)
- **Credits:** 1(2 ECTS)
- **Term:** 2020-2021 Fall
- **Course level:** [MA/MS]
- **Prerequisites:**   Mathematics and Informatics Pre-session for Business Analytics
                       Data Analysis 1: Exploration – Business Analytics track
- **Course drop:** Course can be dropped free of charge 24 hours after the first session. After this date drop is possible until the course is halfway over (late drop fee applies). No changes are allowed past that date.

1. COURSE DESCRIPTION

Data is the new oil." We can hear that in every conference related to data science. This course will teach you how to mine the oil. Every website showed up in our browser is downloadable. Although the data showed in a site has a structure, most of the time the download button does not exists. The data, showed in the browser, is coming from a webserver and if we know the language we can communicate directly with the server. The direct communication makes the data collection very fast and effective. We will download lots of webpages and get familiar some basic operations just like for loop, if statement, lapply, curl, json structure.

2. LEARNING OUTCOMES

By successfully completing the course the students will be able to:

- Understand the basic communication of custom webserver and the browser
- To download any public webpage and process the data on the selected webpage
- Understand the basic of html
- Automate actions on web browser
- To understand the basics of programing.

Other outcomes. The course will also help develop skills in the following areas:

Learning Area	Learning Outcome
Critical Thinking	Look behind the scene (Web browser)
Quantitative Reasoning	
Technology Skills	Advanced R coding

Interpersonal Communication Skills	Working in group
Management Knowledge and Skills	
Cultural Sensitivity and Diversity	
Ethics and Social Responsibility	

3. TEACHING METHODS AND LEARNING ACTIVITIES

The course will be very practical. We will scrape lots of webpages together, and also doing lots of data transformation and data cleaning.

4. ASSESSMENT

Grading will be based on the total score out of 100, in line with CEU's standard grading guidelines

- Home assignment 1 (30%)
- Final project (70%)

5. TECHNICAL/LAPTOP REQUIREMENT

- Laptop
- R and Rstudio installed
- Selectorgadget chrome extension

6. TOPIC OUTLINE AND SCHEDULE

Session	Topics	Readings
1	Demo of the course outcome and starting with R and rvest, understanding Selectorgadget	
2	Structure of html, scraping with R and Rvest	
3	POST and GET request, working with jsonlite	
4	Writing functions to api serveces	
5	Introduction to selenium in python	
6	Alternative tools: postman, Web Scraper extension	

7. SHORT BIO OF THE INSTRUCTOR

Studied physics and IT at university of Pécs. Member of the first-year students of the MSc In Business Analytics 2015. Currently working in Market Épitő Zrt. as a data scientist. Big fan of cryptocurrency, creator of youtube channel https://www.youtube.com/channel/UCWqbZL_js9fh5yUXOiVjmeA . Also made the EU funding (coming to Hungary) transparent and searchable with web scraping and shiny.