**Syllabus**

**Web scraping, from zero to hero**

* **Instructor:** Mihály Orsós (orsosmi@otpbank.hu)
* **Credits:** 1(2 ECTS)
* **Term:**
* **Course level:** [MA/MSc]
* **Prerequisites:**
* **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.

**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. We can think about a website as a csv, if we know what is the data that we want to download. 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.

**Learning outcomes**

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

* Understand the basic communication of custom webserver and the browser
* Look solutions to download any webpage and process the data on the selected webpage
* Understand the basic of html
* Automate actions on web browser

**Assessment (including minimum pass requirement)**

* Choosing a website, download and analyze its data. The analysis should be reproducible.

**Technical/laptop requirement**

Students need to use their own laptops (with R, RStudio, and anaconda python environment) installed during seminars.

**Course schedule and materials for each session**

1. Demo of the course outcome and starting with R and rvest
2. Web scraping with R and rvest
3. Web scraping with POST and GET in R
4. Introduction to selenium
5. Web scraping with Selenium in R and python
6. Postman

**Short bio of the instructor (1 para)**

Studied physics and IT at university of Pécs. Member of the first-year students of the MSc In Business Analytics 2015. Currently work at OTP as an IT developer. Big fan of cryptocurrency, founder of <https://steemdetective.com/> . Also made the EU funding (coming to Hungary) transparent and searchable with web scraping and shiny.