Doctoral School in Economics and Finance

DSEF 2.1. Doctoral Course on *Using the Internet as a Data Source for Social Science – Hands on skills for text mining with Python for economists*

1. Course details

Semester:	Semester 1 - 2020/21
Credit rating:	1 ECTS
Teaching units	15
Pre-requisite(s):	none
Lecturers:	Nikos Askitas (IZA)
Administrator:	Roswitha Glorieux
Tutors:	None
Seminar times and rooms:	please see Point 3
Tutorial times and rooms:	None
Communications	It is important that students should regularly read their University e-mails, as important information will normally be communicated this way.
Mode of assessment:	Assignment
Examination Periods:	TBD
Course WebPage:	Moodle.uni.lu

2. Aims and objectives

The Internet is the place where market activity increasingly and exclusively takes place because digital technology has superior build-in features which make it suitable for optimising matching of supply and demand: from the marriage market, to the transport market, the market for information, the labor market and beyond. On the side of the economic agent ("end user") participation is easy and non-invasive while on the side of the market operator ("back end") it allows for experimentation and complete data recording (hence rewinding and replaying). The Internet is therefore an invaluable trove of data for social scientists. Course participants will acquire concrete skills in python programming for accessing and using that data.

3. Plan of semester

15-17 September 2020

4. Course details (by topics)

The course will start with the legal state of web scraping (covering recent EU copyright legislation and US landmark cases) and proceed to cover the basics of HTML and Python programming in a way tailored to applied economists: HTML will be covered as a "typesetting language" using LaTeX as a comparison (which is a typesetting language familiar to academics) while the Python material will make continuous references to Stata (the most popular language among applied economists).

We will cover Python language constructs (loops, conditionals, functions etc), Python data types (strings, integers, floats etc., lists, dictionaries etc.), basic modules (os, pandas, numpy, etc). We will pay particular attention to re, the regular expression module of Python, which is the workhorse of text mining and we will show the limited (but useful) capabilities of Stata in that area. A variety of skills will also be covered like interacting with the filesystem (reading, writing and updating files), programming a browser (to programmatically surf the internet), building web scraping robots which can log into websites in order to harvest data, downloading data from Google Trends etc.

Everything will be taught by hands on examples. We will use for example RePEc as a source of data to harvest all registered economists, with demographics and other info, the website of the University of Luxembourg to harvest online CV type data, sports sites to download results, the UK parliament website to download petition data, automobile clubs to download traffic jam data etc.

The course material will be written in Jupyter notebooks, which run in a web browser and be made available to every participant. Participants are encouraged to participate with a laptop with Anaconda and the latest python 3 installed and will be assisted in doing so if they have not already. They will then be able to run the scripts in real time during the course and to modify them at will. The collection of all notebooks comprising the course will contain all code snippets necessary to complete the final assignment.

5. Further information about assessment

Examination(s)

Weighting: 100%

Date: tbd

Length:	
Structure:	Pass or Fail. The course will be assessed based on an empirical project applying the techniques covered in the course. The details of the project will be agreed in advance with the lecturer.