Doctoral School in Finance and Economics

Course ID

SELECTED APPLICATION OF WEB-SCRAPING AND TEXT-MINING WITH PYTHON FOR ECONOMISTS. A HANDS-ON APPROACH

1. Course details

Semester:	1
Credit rating:	1 ECTS / 15TU
Pre-requisite(s):	Internet capable Computer with a functioning Anaconda Individual Edition Python 3 installation. All code will be written in Jupyter Notebooks. (https://www.anaconda.com/products/individual)
Lecturer(s):	Nikos Askitas (IZA – Institute of Labor Economics)
Administrator:	Roswitha Glorieux
Tutor(s):	(Melissa Tornari)
Seminar times:	18 May 09:00 - 12:00, 14:00 - 17:00, 19th May 09:00-12:00 20 May 09:00 - 12:00, 14:00 - 17:00,
Room:	TBA
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 Course material: Will be provided on https://cloud.iza.org

2. Aims and objectives

Selected application of web-scraping and text-mining with Python for Economists.
A hands-on approach

Markets are now online and that makes the Internet a prime source of data for social science so web-scraping or using data provisioning APIs (e.g. twitter etc.) is a valuable skill for an economist. Moreover, more and more of such digital data found on the Internet is textual so extracting meaningful quantitative variables from such text (text mining) is equally valuable. Both types of skills enable an economist to create novel datasets for answering old and new questions.

3. Plan of semester

3 classes covering 4 TU each and 1 class covering 3 TU held over a single teaching week.
Classroom, date and time: TBA

4. Course details (by topics)

The course will consist of as many as possible selected applications of mini projects, along the lines of the literature below, demonstrating as large a portion of the complete production pipeline in a research project: Web scrape, clean, text mine, produce quantitative variables, analyze. Topics will include migration, gender inequality, covid 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

Literature Examples in Social Science

- Chaturvedi, S., Mahajan, K., & Siddique, Z. (2021). Words Matter: Gender, Jobs and Applicant Behavior (IZA DP14497). (Preprint: https://docs.iza.org/dp14497.pdf).
- Kim, E., & Patterson, S. (2020). The Pandemic and Gender Inequality in Academia. *Available at SSRN 3666587*. (Preprint: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3666587).
- Nathan, M., Rosso, A., & Bouet, F. (2014). Mapping 'Information Economy'Businesses with Big Data: Findings for the UK. (Preprint: https://ftp.iza.org/dp8662.pdf).
- Xia Z. & Chen J. (2021). Mining The Relationship Between COVID-19 Sentiment and Market Performance. (Preprint: https://arxiv.org/abs/2101.02587v2).
- Cinelli M., Ficcadenti V., & Riccioni J. (2020). The interconnectedness of the economic content in the speeches of the US Presidents. (Preprint: https://arxiv.org/abs/2002.07880)
- Mozafari M., Farahbakhsh R., & Crespi N. (2017). A BERT-Based Learning Approach for Hate Speech Detection in Online social media. (Preprint: https://arxiv.org/pdf/1910.12574.pdf)
- Bana, S. H. (2021). job2vec: Using Language Models to Understand Wage Premia.
- Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2021). What Makes a Good Image? Airbnb Demand Analytics Leveraging Interpretable Image Features. *Management Science*. (Preprint: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2976021).
- Fazel, S., Zhang, L., Javid, B. *et al.* Harnessing Twitter data to survey public attention and attitudes towards COVID-19 vaccines in the UK. *Sci Rep* **11,** 23402 (2021). https://doi.org/10.1038/s41598-021-02710-4

- Abramitzky R., Becker C., Boustan L. P., Card D., Chang S., Jurafsky D., Mendelsohn J., Rashid M., & Voigt R.) (https://legacy.iza.org/en/webcontent/events/izaseminar_description_html?sem_id=3436). Political Speech about Immigration to the US is More Positive but More Polarized Than Any Time in the Past 150 Years (Preprint: Submitted for Publication, No Preprint Available yet).

5. Reference list/ Bibliography

- The Python Language Reference 2022, Python Software Foundation, https://docs.python.org/3/reference/webscarpa.
- Mitchell, R. (2018). *Web scraping with Python: Collecting more data from the modern web*. " O'Reilly Media, Inc.".
- Brooker, P. D. (2019). *Programming with Python for Social Scientists*. Sage.
- Beazley, D., & Jones, B. K. (2013). *Python Cookbook: Recipes for Mastering Python 3*. " O'Reilly Media, Inc.".
- Joshi, Prateek. *Python machine learning cookbook*. Packt Publishing Ltd, 2016.
- McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.".
- Ojeda, T., Bilbro, R., & Bengfort, B. (2018). Applied Text Analysis with Python. Chapter 4. Text Vectorization and Transformation Pipelines.
- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535-74.
- Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press. (Copy: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf)

6. Further information about assessment

Examination(s)	
Weighting:	100%
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.