# GESIS Fall Seminar in Computational Social Science 2021

## Syllabus for week 2:
## "Web Data Collection and Natural Language Processing in Python"


| Lecturers: | Indira Sen | Dr. Arnim Bleier | PhD Roberto Ulloa | Olga Zagovora |
|---|---|---|---|---|
| Email: | indira.sen@gesis.org | arnim.bleier@gesis.org | roberto.uloa@gesis.org | olga.zavora@gesis.org |
| | Mattia Samory<br>Mattia.samory@gesis.org | | | |


Date: September 20-September 24, 2021
Time: 9:00-17:00, with 1 hour lunch break (see day-to-day schedule for details)

## About the Lecturers:

**Indira Sen** is a doctoral candidate at the Computational Social Science Department at GESIS. Her interest lies in understanding biases in inferential studies from digital traces, with a focus on natural language processing. She has experience working with large, unstructured data for social science research.

**Arnim Bleier** is a postdoctoral researcher at the Department Computational Social Science at GESIS. His research interests are in the field of Natural Language Processing and Computational Social Science. In collaboration with social scientists, he develops Bayesian models for the content, structure and dynamics of social phenomena.

**Olga Zagovora** is a doctoral candidate at the Computational Social Science department at GESIS. Prior to joining GESIS, she studied computer science, web and data science. Her research focuses on the evaluation of alternative metrics for measuring scholarly communication and scientific impact. She has experience working with big data for social science research.

**PhD Roberto Ulloa** is a postdoctoral researcher at the Computational Social Science department of GESIS. He researches the role of institutions in shaping societies, and online platforms as forms of digital institutions.

**Mattia Samory** is a postdoctoral researcher at the Computational Social Science department of GESIS. He is currently studying 1) factors in sexist language and its perception; 2) characteristics of the news media landscape online; and 3) open moderation in Reddit.

## Course Description:

Data Science is the interdisciplinary science of extracting interpretable and useful knowledge from potentially large datasets. In contrast to empirical social science, data science methods often serve purposes of exploration and inductive inference. In this course, we aim to provide an introduction on how to tap into the vast amount of digital behavioral data available on Web platforms and processing it to be useful for social science research purposes.

To this end, participants will first learn how to collect data with Web Application Programming Interfaces (APIs) and Web scraping, by employing common Python tools and methods and how to incorporate them into workable data structures. Such APIs will likely include such offered by major social media companies like Reddit, and Youtube (Wikipedia if time permits).

Participants will subsequently be introduced to the basics of Natural Language Processing (NLP) for the analysis of these corpora. As much of the work in NLP is based on Machine Learning (ML), we will begin this section with a basic introduction to ML, followed by an introduction to pre-processing, e.g., data cleaning and feature extraction. We will then cover the application of popular NLP toolkits, some based on simple heuristics and dictionaries, but some also introducing more advanced ML methods.

All course materials will be provided as Python-based Jupyter Notebooks.

## Keywords:

Web Scraping, APIs, Natural Language Processing, Text as data, Social Media, Python

## Course Prerequisites:

- The requirements for attending this course are a functional knowledge of Python and Pandas. We expect the participants to have a working knowledge of Python data structures like lists, dictionaries, and Pandas data frames and how to use these to do basic data wrangling and processing. In case the participants are unfamiliar with these basic concepts of Python programming, we recommend them to attend the course "Introduction to Computational Social Science with Python" in preparation. Additionally, we will also publish a set of external online materials and courses on basic Python and Pandas that participants can use to prepare. The course will include a brief refresher on the basics of Python and Pandas in the beginning – this does however not replace a proper introduction into Python.
- Some previous knowledge of statistics would be beneficial, although not mandatory.
- Participants have preferably worked in a Jupyter Notebook environment before. Detailed installation instructions on how to access the Jupyter Notebook cloud environment will be provided before the start of the course.

## Target Group:

Participants will find the course useful if:

- They are interested in obtaining digital behavioral data from Web platforms through different APIs and Web Scraping.
- They need to structure textual user contributions to study social phenomena.
- They are interested in learning the basics of applying some Natural Language Processing, including basic Machine Learning applications.

We expect this tutorial to be of interest for participants from a variety of disciplinary backgrounds (e.g., Economics, linguistics, sociology, psychology, political science, demography), particularly those who are interested in leveraging novel forms of digital traces for drawing inferences.

## Course and Learning Objectives:

Participants will obtain a working knowledge of how web and social media is collected through a detailed introduction to Web APIs and Web scraping and corresponding tools. Participants will obtain knowledge about typical data types and structures encountered when dealing with digital behavioral data from the Web, and how to apply selected NLP methods and tools in Python to structure natural language texts; and they will learn how this approach differs from those typically encountered in survey-based or experimental research. This will enable them to identify benefits and pitfalls of these data types and methods in their field of interest and will thus allow them to select and appropriately apply the covered NLP methods to large datasets in their own research. The knowledge obtained in this course provides a starting point for participants to investigate specialized methods for their individual research projects.

## Organisational Structure of the Course:

The course will be structured based on different subthemes of Web data collection, working with digital human traces from the web and processing for analysis. Lectures will be interactive, and the use of Jupyter Notebooks allows participants to reproduce the steps along the research pipeline while we introduce the topics. Each lecture will be a combination of conceptual sections and hands-on programming examples. Additionally, participants will have the opportunity to cement and test their understanding of different concepts in regular exercise and feedback rounds, where instructors provide support, advice, and troubleshooting.

## Software requirements:

Participants should have an installation of Anaconda ready, along with Python 3.7+, and Jupyter Notebooks. Anaconda is an open data science platform powered by Python which can be downloaded here: https://www.anaconda.com/products/individual . It comes with many code libraries / packages for Python already installed. It also comes equipped with Jupyter Notebooks. We will be working with Python 3.7 and we will use Jupyter Notebooks for the exercises. While we plan to work in Jupyter Notebooks in our ready-to-go cloud-based environment notebooks.gesis.org, local installations of Anaconda are needed in rare cases of downtime of this service.

## Recommended Literature to look at in advance:

- *Web Scraping with Python*, 2nd Edition by Ryan Mitchell, April 2018, O'Reilly Media, Inc., ISBN: 9781491985571
- "Introduction to Machine Learning with Python" by Andreas Mueller https://github.com/amueller/introduction_to_ml_with_python
- "Python for Data Analysis" by Wes McKinney https://github.com/wesm/pydata-book
- “A Code-First Introduction to Natural Language Processing” by Rachel Thomas https://github.com/fastai/course-nlp

## Day-to-day schedule and literature:

**Day 1: Introduction and Understanding APIs**

9:00 Welcome
9:10 Course intro & Overview
9:50 Recap Pandas and Data Wrangling
11:05 Break
11:20 Lecture: Web Data Acquisition Introduction
11:55 Lecture: Understanding APIs 1 (w/ abgeordnetenwatch.de)
12:55 Lunch
13:55 Exercise: Understanding APIs 1
14:30 Lecture: Understanding APIs 2
15:00 Break
15:15 Exercise: Understanding APIs 2
15:45 Reddit: intro to the API + data collection: posts
17:00 End

*Literature:* Chapter 12 of *Web Scraping with Python*, 2nd Edition by Ryan Mitchell, April 2018, O'Reilly Media, Inc., ISBN: 9781491985571

**Day 2: Working with the Reddit and YouTube APIs**

9:00 Reddit Exercise 1
9:30 Reddit data collection 2: comments and users
10:30 Break
10:45 Reddit Data Wrangling and Cleaning
11:45 Lunch
12:45 Reddit Exercise 2
13:15 The YouTube Data API
14:15 Break
14:30 Collecting YouTube Data
15:45 Preprocessing YouTube Comments
17:00 End

*Literature:* -

**Day 3: Webscraping**

9:00 Youtube Exercise
9:45 Lecture: Webscraping 1
11:15 Exercise: Webscraping 1
12:00 Lunch
13:00 Lecture: Webscraping 2
14:30 Break
14:45 Exercise: Webscraping 2
15:30 Q & A / Repetition / Mini projects
17:00 End

*Literature:* Chapters 1,2,7,8 of *Web Scraping with Python*, 2nd Edition by Ryan Mitchell, April 2018, O'Reilly Media, Inc., ISBN: 9781491985571

**Day 4: Machine Learning and Text as Data**

09:00 Introduction to Supervised Learning
10:15 Break
10:30 Hands-on Scikit-learn
12:00 Lunch
13:00 Unsupervised learning
14:15 Break
14:30 Exercise
15:30 Break
15:45 Text as Data
17:00 End

*Literature:*

"Elegant SciPy" by Juan Nunez-Iglesias et al. https://github.com/elegant-scipy/elegant-scipy

1. "Introduction to Machine Learning with Python" by Andreas Mueller https://github.com/amueller/introduction_to_ml_with_python
2. "Python for Data Analysis" by Wes McKinney https://github.com/wesm/pydata-book

**Day 5: Applications of Natural Language Processing**

09:00: Finding Topics in Text
10:15: Break
10:30: Deep Learning for NLP
12:00: Lunch
13:00: Finding the right Model
14:15: Break
14:30: Exercise
15:45: Summary, evaluation, certificates

*Literature:*

1. "A Code-First Introduction to Natural Language Processing" by Rachel Thomas https://github.com/fastai/course-nlp
2. "Natural Language Processing" by Jacob Eisenstein https://github.com/jacobeisenstein/gt-nlp-class/tree/master/notes