GESIS Fall Seminar in Computational Social Science 2021

Syllabus for week 3:
“A practical introduction to machine learning in Python"

Lecturers:	Dr. Damian Trilling	Dr. Anne Kroon
Email:	d.c.trilling@uva.nl	a.c.kroon@uva.nl

Date: September 27-October 1, 2021
Time: 9:00-17:00, including a lunch break from about 12:30-13:30 and several coffee breaks

About the Lecturers:

Damian Trilling is an associate professor in political communication at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam. He uses computational methods to study today’s news media landscape, and how people interact with the news in an era of social network sites and recommender systems.

Anne Kroon is an assistant professor in corporate communication at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam. She draws on NLP techniques to study bias and stereotypes in media coverage of diverse social groups.

Course Description:

The course will provide insights into the concepts, challenges and opportunities associated with data so large that traditional research methods (like manual coding) cannot be applied anymore and traditional inferential statistics start to lose their meaning. Participants are introduced to strategies and techniques for capturing and analyzing digital data in communication contexts using Python. The course offers hands-on instructions regarding the several stages of computer-aided content analysis. More in particular, students will be familiarized with preprocessing methods, analysis strategies and the visualization and presentation of findings. The focus will be in particular on Machine Learning techniques to analyze quantitative textual data, amongst which both deductive (e.g., supervised machine learning and inductive (e.g., unsupervised machine learning) approaches will be discussed.

To participate in this course, students are expected to be interested in learning how to write own programs where off-the-shelf software is not available.

Keywords:

Python, Supervised Machine Learning, Unsupervised Machine Learning

Course Prerequisites:

- Knowledge of basic statistics (linear and logistic regression)
- Some experience with computational methods, programming in general, and/or statistical languages (but not necessarily Python) is highly recommended to participate in this course. During the first day of the course, we will discuss some fundamental aspects of coding in Python at a fast pace. In order to follow along, we recommend those who have no previous experience with computational methods or statistical languages to take part in the course: Introduction to CSS with Python (week 1).
- Participants are expected to have a working Python environment installed (see below), and we strongly recommend that participants spend a couple of hours with one of the many free only resources to familiarize themselves with the very basics of Python to have an easier start.

Target Group:

Participants will find the course useful if:

- Are social scientists who have the ambition to model quantitative textual data. Specifically, those who aim to describe, explain or predict the content of large-scale textual data using computation techniques are likely to benefit from participating in this course.
- Note that non-textual data, such as images or networks, are not at the center of this course. Techniques we cover are partly generalizable to such types of data, but note that the course is not tailored towards them.

Course and Learning Objectives:

By the end of the course participants will:

- be able to identify research methods from computer science and computational linguistics which can be used for research in the domain of social science;
- understanding of the principles of supervised and unsupervised machine learning;
- be able to explain the principles of these methods and describe the value of these methods;
- know to analyze textual data;
- have basic knowledge of the programming language Python and know how to use Python-modules for questions relevant in the domain of the social sciences;
- be able to independently analyze quantitative textual data using machine learning techniques.

Organisational Structure of the Course:

In the morning, we will have three hours of online lectures, in which we will explain the topic of the day both from a theoretical-conceptual point of view as well as from a practical point of view (i.e., walking you through code examples). We may have small in-class exercises in between, if necessary.

In the afternoon, students work on larger exercises in which they implement the techniques we covered. We provide example datasets, but it is also possible (and encouraged) to try to apply the techniques to own datasets. Due to the online setting, participants can either opt to work on their own (and just re-join for plenary feedback moments), or try to solve problems together with one or multiple classmates in a breakout room. Lecturers will provide feedback on the (attempted) solutions of participants, and also provide example solutions.

Software requirements:

Participants need to have a current Python environment installed and need to be able to install and update packages on their own. All relatively recent versions of Python (in general, 3.6 or higher) should be fine. If you still have an older version, you may not be able to run the example code 1:1 but need to adapt it. Make sure you have recent versions of crucial packages such as pandas, numpy, scipy, scikit-learn, gensim, and keras installed. If in doubt, check how to update them. One option to achieve all of this is to simply install the newest version of the so-called Anaconda distribution, even though this is by no means necessary (in fact, both of us usually install our packages by hand instead of using Anaconda).

Long Course Description:

Social scientists are more and more confronted with the analysis of large scale datasets. Often, these are data from online sources, and often, they contain some form of textual data. One can think of data from social media, but also large archives. Often, this development is referred to as a move towards “computational social science”.

This course introduces you to automated content analysis of data that typically comes in amounts that are too voluminous for manual coding and for traditional point-and-click applications: tweets, blogposts, articles from RSS-feeds, etc. We will use the programming language Python, which is very flexible and highly suitable for this end. It also scales very nicely---meaning that you can use it now for some smaller projects, but it can also be used on immense data sets.

The course will focus particularly on the modelling of textual data using machine learning techniques. Unsupervised Machine Learning techniques are useful to offer a descriptive, inductive account of the data at hand. For example, topic modelling might be useful when one wants to understand which topics were most prominently discussed on online news sites. Supervised Machine Learning, on the other hand, is useful when has a clear preconception of the concepts one wishes to measure in textual data. Using annotated (i.e., coded) data, we will dive into the process of 'teaching' the computer to code unseen data.

The techniques we cover can also be used to analyze non-textual data. However, note that the course focuses on textual data. We will provide example datasets, but you are also encouraged to bring your own.

Also note that we are taking an applied approach here. While we will discuss the inner workings of different models to the extent that this helps our understanding of the workings of the method, we will neither go into all mathematical details nor focus too extensively on the internal implementation of the algorithms we use.

Recommended Literature to look at in advance:

Boumans, J. W. & Trilling, D. (2016). Taking stock of the toolkit. An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism, 4*, 8-23. doi: 10.1080/21670811.2015.1096598.

van Atteveldt, W., Trilling, D., Arcila Calderon, C. (in press): Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.
The book is currently in press. The participants will be provided with full-text access.

Day-to-day Schedule and Literature:

Day	Topic(s)
1	**Introduction and Getting Started in Python** ▪ Introduction and overview of the course ▪ Principles of quantitative textual analysis for social scientists ▪ Getting started with programming in Python. Introduction to the main concepts (such as data types, functions and methods) ▪ Practical discussion of benefits and drawbacks of working with different IDEs, as well as working with specific modules (such as pandas) versus native Python data structures.
	Suggested reading: ▪ Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1* (1), 1–12. doi: 10.1177/2053951714528481 ▪ van Atteveldt, W., Trilling, D., Arcila Calderon, C. (in press): Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R. *Chapters 1-4*
2	**Preparing for Analysis: From text to features** ▪ Introduction to the toolkit accessible to social scientists working with 'big' textual datasets ▪ Inductive and deductive approaches to computer-aided content analysis ▪ Exploratory techniques to explore your data ▪ When, why, and how do we pre-process? ▪ Regular expressions ▪ Natural Language Processing with NLTK and spacy ▪ From text to feature: count vectorizers and tf-idf vectorizers
	Suggested reading: ▪ Boumans, J. W. & Trilling, D. (2016). Taking stock of the toolkit. An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. Digital Journalism, 1, 8-23. doi: 10.1080/21670811.2015.1096598 ▪ Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189. doi:10.1017/pan.2017.44 ▪ van Atteveldt, W., Trilling, D., Arcila Calderon, C. (in press): Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R. *Chapters 9-10*
3	**Unsupervised machine learning** ▪ Principles and techniques of Unsupervised Machine Learning techniques; ▪ Principal Component Analysis, k-means clustering, and hierarchical clustering ▪ Topic modelling with Latent Dirichlet Allocation (LDA) ▪ Hands-on instructions to apply these techniques, using modules such as scikit-learn and gensim
	Suggested reading: ▪ Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., . . . Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12* (2-3), 93–118. doi: 10.1080/19312458.2018.1430754 ▪ van Atteveldt, W., Trilling, D., Arcila Calderon, C. (in press): Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R. *Chapters 7.3 and 11.5*
4	**Supervised machine learning** ▪ Principles and techniques of Supervised Machine Learning;

	▪ Discussion of how logistic regression and Naive Bayes classifiers can be used to predict, for instance, movie ratings or topics of news articles. ▪ Evaluation metrics (accuracy, precision, recall, ...) ▪ Hands-on instructions to apply these techniques, using modules such as scikit-learn
	<u>Suggested reading:</u> ▪ van Atteveldt, W., Trilling, D., Arcila Calderon, C. (in press): Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R. *Chapters 8 (except 8.4) and 1*
5	**Finding the best model and communicating results** ▪ Alternative models (e.g., Random Forests) → DAG 4 ▪ Advanced Supervised Machine Learning (e.g., cross-validation, model selection and tuning) ▪ Visualization and presentation of findings ▪ Outlook: Recent developments that are out of scope of this course (e.g. deep learning with keras) →uUITREIDEN
	<u>Suggested reading:</u> ▪ van Atteveldt, W., Trilling, D., Arcila Calderon, C. (in press): Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R. *Chapter 8.5*