GESIS Fall Seminar in Computational Social Science 2023

Syllabus for week 3:
"Introduction to Machine Learning for Text Analysis with Python"

Lecturers:	Damian Trilling	Anne Kroon
Affiliation:	University of Amsterdam	University of Amsterdam
Email:	d.c.trilling@uva.nl	a.c.kroon@uva.nl

Date: September 25-29, 2023
Time: 10:00-17:00

About the Lecturers

Damian Trilling is Associate Professor *Communication in the Digital Society* at the Department of Communication Science, University of Amsterdam, where he is a member of the Programme Group *Political Communication and Journalism* and affiliated with the Amsterdam School of Communication Research (ASCoR). Damian is interested in studying news flows in the contemporary online media ecosystem, for which he uses sometimes traditional, but mostly computational methods. He is one of the authors of the textbook *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R,* and funding associate editor of the journal *Computational Communication Research.*

Anne Kroon is Assistant Professor *Corporate Communication* at the Department of Communication Science, University of Amsterdam. Anne's research primarily focuses on the role of algorithms in recruitment and hiring as a means to address bias, as well as (biased) presentation of minorities in media content. Both in her teaching and in her research, computational methods of (text) analysis take center stage.

Course Description

The course will provide insights into the concepts, challenges and opportunities associated with data so large that traditional research methods (like manual content analysis) 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 pre-processing 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.
This is a beginner's course. Participants who are looking to learn about the latest developments in machine learning for textual data (such as transformer models) should consider taking a different course, e.g. "From Embeddings to Transformers: Advanced Text Analysis for Social Scientists". These techniques will be (briefly) discussed towards the end of the course, but the focus lies on the basics of natural language processing and classical machine learning in Python.

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 little previous experience with computational methods or statistical languages to take part in the course "Introduction to CSS with Python" (week 1, 11-15 September).
- 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 online resources to familiarize themselves with the very basics of Python to have an easier start. For a basic introduction or refresher to Python programming, participants may also consider taking the online workshop "Introduction to Python" that takes place from 04-06 September 2023.

Target Group

Participants will find the course useful if:

- They 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. Participants interested in working with images or networks might be interested in one of the following two courses: "Automated Image and Video Data Analysis with Python" in Week 2 (18-22 September) or "Social Network Analysis with R" in Week 3 (25-29 September).

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
- have an 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 how 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

Organizational Structure of the Course

In the morning, we will have 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. Participants can either opt to work on their own or try to solve problems together with one or multiple classmates. Lecturers will provide feedback on the (attempted) solutions of participants, and also provide example solutions.

Software and Hardware 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, 3.8 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). Additionally, it is advisable to have access to Google Colab. Therefore, please ensure that you have a Google account and can execute code through Google Colab.

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. (2022*): Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R*. Hoboken, NJ: Wiley. The printed version of this book can be ordered online. The full text is freely available at https://cssbook.net/

Day-to-day Schedule and Literature

Day 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.
- Conducting an exercise that focuses on setting up our first simple machine learning classifier.

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. (2022): van Atteveldt, W., Trilling, D., Arcila Calderon, C. (2022): *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* Hoboken, NJ: Wiley. *Chapters 1-4.* https://cssbook.net/

Day 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 features: 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. (2022): van Atteveldt, W., Trilling, D., Arcila Calderon, C. (2022): *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* Hoboken, NJ: Wiley. *Chapters 9-10.* https://cssbook.net/

Day 3: Unsupervised Machine Learning

- Principles and techniques of Unsupervised Machine Learning techniques
    - e.g., a brief introduction to Principal Component Analysis, k-means clustering, and hierarchical clustering
- Topic modeling with Latent Dirichlet Allocation (LDA)
- Hands-on instructions to apply these techniques, using modules such as scikit-learn and gensim
- Comparing techniques of unsupervised learning with supervised learning

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. (2022): van Atteveldt, W., Trilling, D., Arcila Calderon, C. (2022): *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* Hoboken, NJ: Wiley. *Chapters 7.3 and 11.5.* https://cssbook.net/

Day 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
- Alternative models (e.g., Random Forests)
- Advanced Supervised Machine Learning (e.g., cross-validation, grid search, model selection, and tuning)

Suggested reading:

van Atteveldt, W., Trilling, D., Arcila Calderon, C. (2022): *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* Hoboken, NJ: Wiley. *Chapters 8 (except 8.4) and 1.* https://cssbook.net/

Day 5: Recent Developments in Machine Learning

- Visualization and presentation of findings
- Outlook: Recent developments that are out of the scope of this course (e.g., embedding models, Transformer models, deep learning with keras)

Suggested reading:
van Atteveldt, W., Trilling, D., Arcila Calderon, C. (2022): *Computational Analysis of Communication: A practical introduction to the analysis of texts, networks, and images with code examples in Python and R.* Hoboken, NJ: Wiley. *Chapter 8.5.* https://cssbook.net/