Big Data and Automated Content Analysis
Part I+II (12 ECTS)

Cursusdossier

dr. Damian Trilling

Graduate School of Communication
University of Amsterdam

d.c.trilling@uva.nl
www.damiantrilling.net
@damian0604

Office: REC-C, 8th floor

Academic Year 2022/23

Contents

Chapter 1

Short description of the course

"Big data" refers to data that are more voluminous, but often also more unstructured and dynamic, than traditionally the case. In Communication Science and the Social Sciences more broadly, this in particular concerns research that draws on Internet-based data sources such as social media, large digital archives, and public comments to news and products. This emerging field of studies is also called *Computational Social Science* (Lazer et al., 2009) or, narrowed down to the analysis of communication, *Computational Communication Science* (Shah, Cappella, & Neuman, 2015).

The course will provide insights in the concepts, challenges and opportunities associated with data so large that traditional research methods (like manual coding) cannot be applied any more and traditional inferential statistics start to loose their meaning. Participants are introduced to strategies and techniques for capturing and analyzing digital data in communication contexts. We will focus on (a) data harvesting, storage, and preprocessing and (b) computer-aided content analysis, including natural language processing (NLP) and computational social science approaches. In particular, we will use advanced machine learning approaches and models like word embeddings.

To participate in this course, students are expected to be interested in learning how to write own programs in Python. Some basic understanding of programming languages is helpful, but not necessary to enter the course. Students without such knowledge are encouraged to follow one of the many (free) online introductions to Python to prepare.

Chapter 2

Exit qualifications

(Note: In this chapter"advanced research designs and methods" in the following refers to techniques of computational communication science as covered in Part I (basic text processing, data retrieval from web sources) and Part II (supervised machine laring, and unsupervised machine learning) of this course).

The course contributes to the following three exit qualifications of the Research Master in Communication Science:

Expertise in empirical research

3. Knowledge and Understanding: Have in-depth knowledge and a thorough understanding of advanced research designs and methods

4. Skills and abilities: Are able, independently and on their own, to set up, conduct, report and interpret advanced academic research.

Academic abilities and attitudes

6. Attitude: Accept that scientific knowledge is always 'work in progress' and that arguments must be considered and conclusions drawn on the basis of empirical results and valid criticism.

The exit qualifications are elaborated in the following 11 specifications: 3. Knowledge and Understanding: Have in-depth knowledge and a thorough understanding of advanced research designs and methods.

3.1. Have in-depth knowledge and a thorough understanding of advanced research designs and methods, including their value and limitations.

3.2. Have in-depth knowledge and a thorough understanding of advanced techniques for data analysis.

4. Skills and abilities: Are able, independently and on their own, to set up, conduct, report and interpret advanced academic research.

4.1 Are able to formulate research questions and hypotheses for advanced empirical studies

4.2 Are able to develop a research plan, choose appropriate and suitable research designs and methods for advanced empirical studies, and justify the underlying choices.

4.3 Are able to assess the validity and reliability of advanced empirical research, and to judge the scientific and professional value of findings from advanced empirical research.

4.4 Are able to apply advanced empirical research methods.

6. Academic attitudes

6.1 Regularly asses their own assumptions, strengths and weaknesses critically.

6.2 Accept that scientific knowledge is always 'work in progress' and that something regarded as 'true' may be proven to be false, and vice-versa.

6.3 Are keen to acquire new knowledge, skills and abilities.

6.4 Are willing to share and discuss arguments, results and conclusions, including submitting one's own work to peer review.

6.5 Are convinced that academic debates should not be conducted on the basis of rhetorical qualities but that arguments must be considered and conclusions drawn on the basis of empirical results and valid criticism.

Chapter 3

Testable objectives

3. Knowledge and Understanding: Have in-depth knowledge and a thorough understanding of advanced research designs and methods.

3.1. Have in-depth knowledge and a thorough understanding of advanced research designs and methods, including their value and limitations.

3.2. Have in-depth knowledge and a thorough understanding of advanced techniques for data analysis.

A Students can explain the research designs and methods employed in existing research articles on Big Data and automated content analysis.

B Students can on their own and in own words critically discuss the pros and cons of research designs and methods employed in existing research articles on Big Data and automated content analysis; they can, based on this, give a critical evaluation of the methods and, where relevant, give advice to improve the study in question.

C Students can identify research methods from computer science and computer linguistics which can be used for research in the domain of communication science; they can explain the principles of these methods and describe the value of these methods for communication science research.4. Skills and abilities: Are able, independently and on their own, to set up, conduct, report and interpret advanced academic research.

4.1 Are able to formulate research questions and hypotheses for advanced empirical studies

4.2 Are able to develop a research plan, choose appropriate and suitable research designs and methods for advanced empirical studies, and justify the underlying choices.

4.3 Are able to assess the validity and reliability of advanced empirical research, and to judge the scientific and professional value of findings from advanced empirical research.

4.4 Are able to apply advanced empirical research methods.

D Students can on their own formulate a research question and hypotheses for own empirical research in the domain of Big Data.

E Students can on their own chose, execute and report on advanced research methods in the domain of Big Data and automatic content analysis.

F Students know how to collect data with scrapers, crawlers and APIs; they know how to analyze these data and to this end, they have basic knowledge of the programming language Python and know how to use Python-modules for communication science research.

6. Academic attitudes

6.1 Regularly asses their own assumptions, strengths and weaknesses critically.

6.2 Accept that scientific knowledge is always 'work in progress' and that something regarded as 'true' may be proven to be false, and vice-versa.

6.3 Are keen to acquire new knowledge, skills and abilities.

6.4 Are willing to share and discuss arguments, results and conclusions, including submitting one's own work to peer review.

6.5 Are convinced that academic debates should not be conducted on the basis of rhetorical qualities but that arguments must be considered and conclusions drawn on the basis of empirical results and valid criticism.

G Students can critically discuss strong and weak points of their own research and suggest improvements.

H Students participate actively: reading the literature carefully and on time, completing assignments carefully and on time, active participation in discussions, and giving feedback on the work of fellow students give evidence of this.

Chapter 4

Planning of testing and teaching

The seminar consists of 28 meetings, two per week. Each week, in the first meeting, the instructor will give short lectures on the key aspects of the week, followed by seminar-style discussions. Theoretical considerations regarding Big Data and Automated Content Analysis are discussed, and techniques for analyzing Big Data are presented. We also discuss examples from the literature, in which these techniques are applied.

The second meetings each week are practicum-meetings, in which the students will apply what the techniques they have learned to own data sets. Here, they can also deepen their understanding of software tools, prepare their projects and get hands-on help. While there are in-class assignments as well as occasional assignments for at home (e.g., completing an online-tutorial to prepare for class), these are not graded.

To complete the course, next to active participation, the students have to successfully complete three summative graded assignments: two mid-term take-home exam and an individual project, in which they derive an empirical question from a theoretical starting point, and then do an Automated Content Analysis to answer the question. See Chapter 7 for details.

Chapter 5

Literature

The following schedule gives an overview of the topics covered each week, the obligatory literature that has to be studied each week, and other tasks the students have to complete in preparation of the class. In particular, the schedule shows which chapter of van Atteveldt, Trilling, and Arcila Calderón (2022) will be dealt with. Note that some basic chapters that explain how to install the software we are going to use have to be read before the course starts.

Next to the obligatory literature, the following books provide the interested student with more and deeper information. They are intended for the advanced reader and might be useful for final individual projects, but are by no means required literature. Bear in mind, though, that you may encounter slightly outdated examples.

- VanderPlas, 2016: A book on numpy, pandas, scikit-learn and more. It can also be read online for free on `https://jakevdp.github.io/PythonDataScienceHandbook/`, and the contents are avaibale as Jupyter Notebooks as well `https://github.com/jakevdp/PythonDataScienceHandbook`.
- The pandas cookbook by Julia Evans, a collection of notebooks on github: `https://github.com/jvns/pandas-cookbook`.
- Hovy, 2020: A thin book on bottom-up text analysis in Python with both a bit more math background and ready-to-use Python code implementations.
- Salganik, 2017: Not a book on Python, but on research methods in the digital age. Very readable, and a lots of inspiration and background about techniques covered in our course.

Chapter 6

Specific course timetable

Before the course starts: Prepare your computer.

✔ CHAPTER 1: INTRODUCTION
Make sure that you have a working Python environment installed on your computer. You cannot start the course if you have not done so.

PART I: Basics of Python and ACA

Week 1: Programming for Computational (Communication|Social) Scientists

Wednesday, 8–2. Lecture with exercises.

We discuss what Big Data and Computational (Social|Communication) Science are. We talk about challenges and opportunities as well as the implications for the social sciences in general and communication science in particular. We also pay attention to the tools used in CSS, in particular to the use of Python.

Mandatory readings (in advance): Boyd and Crawford (2012), Kitchin (2014), Hilbert et al. (2019).

Additionally, the journal *Commmunication Methods and Measures* had a special issue (volume 12, issue 2–3) about Computational Communication Science. Read at least the editorial (van Atteveldt & Peng, 2018), but preferably, also some of the articles (you can also do that later in the course).

Towards the end of the lecture, we will make first contact with writing code.

Friday, 1–2. Lecture with exercises.

✔ Chapter 3: Programming concepts for data analysis
✔ Chapter 4: How to write code

You will get a very gentle introduction to computer programming. During the lecture, you are encouraged to follow the examples on your own laptop.

We will do our first real steps in Python and do some exercises to get the feeling with writing code.

Week 2: From files and APIs to lists, dictionaries, or dataframes

✔ Chapter 5: From file to dataframe and back

We talk about file formats such as `csv` and `json`; about encodings; about reading these formats into basic Python structures such as dictionaries and lists as opposed to reading them into dataframes; and about retrieving such data from local files, as parts of packages, and via an API.

Wednesday, 15–2. Lecture.

A conceptual overview of different file formats and data sources, and some practical guidance on how to handle such data in basic Python and in Pandas.

Friday, 17–2. Lab session.

We will exercise with the data structures we learned in week 1, as well as with different file formats.

Week 3: Data wrangling and exploratory data analysis

Of course, you don't need Python to do statistics. Whether it's R, Stata, or SPSS – you probably already have a tool that you are comfortable with. But you also do not want to switch to a different environment just for getting a correlation. And you definitly don't want to do advanced data wrangling in SPSS... This week, we will discuss different ways of organizing your data (e.g., long vs wide formats) as well as how to do conventional statistical tests and simple plots in Python.

Wedneday, 22–2. Lecture.

✔ Chapter 6: Data wrangling
✔ Chapter 7.1. Simple exploratory data analysis
✔ Chapter 7.2. Visualizing data

We will learn how to get your data in the right shape and how to get a first understanding of your data, using exploratory analysis and visualization techniques. We will cover data wrangling with pandas: converting between wide and long formats (melting and pivoting), aggregating data, joining datasets, and so on.

Friday, 24–2. Lab session.

We will apply the techniques discussed during the lectures to multiple datasets.

Week 4: Machine learning basics

In this week, we will make the transition from classic statistical modeling as you know it from your previous courses to machine learning. We will discuss how both approaches are related (or even identical) and where the differences are.

Wednesday, 1–3. Lecture

✔ Chapter 7.3. Clustering and Dimensionality Reduction
✔ Chapter 8: Statistical Modeling and Supervised Machine Learning
✖ (you can skip 8.4 Deep Learning for now)

We will discuss what unsupervised and supervised machine learning are, what they can be used for, and how they can be evaluated.

Friday, 3–3. Lab session.

Departuring from a brief encounter with statsmodels (Seabold & Perktold, 2010), a library for statistical modelling, you will learn how to work with scikit-learn (Pedregosa et al., 2011), one of the most well-known machine learning libraries.

Week 5: Processing textual data

In this week, we will dive into how to deal with textual data. How is text represented, how can we process it, and how can we extract useful information from it? Unfortunately, text as written by humans usually is pretty messy. We will therefore dive into ways to represent text in a clean(er) way. We will introduce the Bag-of-Words (BOW) representation and show multiple ways of transforming text into matrices.

Wedneday, 8–3. Lecture.

✔ Chapter 9: Processing text
✔ Chapter 10: Text as data
✔ Chapter 11, Sections 11.1–11.3: Automatic analysis of text

This lecture will introduce you to techniques and concepts like lemmatization, stopword removal, n-grams, word counts and word co-occurrances, and regular expressions. We then proceed to introducing BOW representations of text.

Additional recommended background reading on stopwords: Nothman, Qin, and Yurchak (2018).

Friday, 10–3. Lab session.

You will combine the techiques discussed on Wednesday and write a first automated content analysis script.

Take-home exam

In week 5, the first midterm take-home exam is distributed after the Friday meeting. The answer sheets and all files have to be handed in no later than the day before the next meeting, i.e. Tuesday evening (14–3, 23.59).

Week 6: Supervised Approaches to Text Analysis

✔ Chapter 11, Section 11.4: Automatic analysis of text

Wednesday, 15–3. Lecture.

We discuss why and when to choose supervised machine learning approaches as opposed to dictionary- or rule-based approaches, and explore how BOW representations can be used as an input for supervised machine learning.

Mandatory reading: Boumans and Trilling (2016).

Friday, 17–3. Lab session.

Exercises with scikit-learn.

Week 7: Supervised Approaches to Text Analysis II

Wednesday, 22–3. Lecture.

We will continue with the topic in week 8, with special attention on how to find the best model using techniques such as crossvalidation and gridsearch.

Friday, 24–3. Lab session.

Exercises with scikit-learn.

Break between block 1 and 2

PART II: Advanced analyses

Week 8: Beyond Bag-of-Words

Wednesday, 5–4. Lecture with exercises.

✔ Chapter 10.3.3. Word Embeddings
✔ Chapter 8.3.5. Neural Networks

✔ Chapter 8.4. Deep Learning

In this week, we will talk about a problem of standard forms of ACA: they treat words as independent from each other, and as either present or absent. For instance, if "teacher" is a feature in a specific model, and a text mentions "instructor", then this is not captured – even though it probably should matter, at least to some extend. Word embeddings are a technique to overcome this problem. But also, they can reveal hidden biases in the texts they are trained on. You will also be provided with examples for how to apply a word2vec model and get a short introduction to keras.

Mandatory readings (in advance): Kusner, Sun, Kolkin, and Weinberger (2015) and Garg, Schiebinger, Jurafsky, and Zou (2018).

Friday, 8–4. No meeting (Good Friday).

Week 9: Transformers

Wednesday, 12–4.

In this lecture, we will introduce Transformer models such as BERT. These models have revolutionized the field in many ways. On the one hand, they have lead to large perfomance increases for many tasks, and they make impressive applications like ChatGPT possible. On the other hand, they form a black box and require extraordinary resources to create. We will discuss the idea behind transformers, introduce the concept of *finetuning* such a pre-trained model, and briefly mention few-shot and zero-shot learning.

Mandatory reading (in advance): Lin, Welbers, Vermeer, and Trilling (2023) and Bender, Gebru, McMillan-Major, and Shmitchell (2021).

Friday, 14–4. Lab session.

We will exercise with finetuning a transformer model using the Huggingface library.

Week 10: Intermezzo: How to gather online data

By now, you know a lot about the analysis of existing data sets – but the big elephant in the room is, of course: how can you get the data to answer your specific research questions?

Reserve some time for exercising in this week. Web scraping can be really hard, because there are so many specifics of specific websites to consider. After all, every website is different, and we need to customize scrapers for every site! At the same time, it is one of the most useful techniques to know, and the majority of students in previous cohorts used web scraping as a part of their final project.

Wednesday, 19–4. Lecture.

✔ Chapter 12: APIs and web scraping

We first discuss the principles behind so-called application programming interfaces (APIs) and learn how to use them to retrieve JSON data. However, not all data that we may be interested in are available in such a format. Therefore, we move on to explore techniques to download data from web pages and to extract meaningful information like the text (or a photo, or a headline, or the author) from an article on `http://nu.nl`, a review (or a price, or a link) from `http://kieskeurig.nl`, or similar.

Friday, 21-4. Lab session.

Exercises with APIs and/or web scraping.

Take-home exam

In week 10, the second midterm take-home exam is distributed after the Friday meeting. The answer sheets and all files have to be handed in no later than the day before the next meeting, i.e. Tuesday evening (25–4, 23.59).

Week 11: Unsupervised Machine Learning for Text

✔ Chapter 11.5. Unsupervised text analysis: Topic modeling and beyond

Wednesday, 26-4. Lecture.

In Part I of this course, we introduced the fundamental distinction between supervised and unsupervised machine learning. Also, when talking about embeddings and transformers, the idea of the unsupervised training on large corpora of text came up again. What we did not discuss so far is the use of unsupervised models for the explorative analysis of text.

A first approach that has historically been employed to do this is to simply apply unsupervised methods such as PCA and k-means clustering on a BOW representation of text – something that you could actually have done already with your knowledge from Part I. Starting from there, we proceed to discuss a second approach, Latent Dirichlet Allication (LDA), also referred to as (a form of) topic modeling.

Both approaches have been influential for the field, but are less of a silver bullet then many students and researchers seem to think. We will therefore introduce a much more state-of-the-art approach that is build on top of a pre-trained Transformer instead of relying on a BOW representation.

Mandatory readings (in advance): Maier et al. (2018) and Grootendorst (2022)

Friday, 28–4. Mo meeting - day after Koningsdag.

Week 12: No Teaching (UvA Teaching Free Week)

Week 13: Multimedia data

Wednesday, 10-5. Lecture

✔ Chapter 14 Multimedia data
We will look beyond text and discuss approaches to the computational analysis of multimedia data.

Friday, 12-5. Lab Session

Opportunity to exercise with APIs and libraries presented during the lecture and/or previous week.

Week 14: Wrapping up

Wendesday, 17–5. Open Lab.

Open meeting with the possibility to ask last (!) questions regarding the final project.

Friday, 20–5. No meeting - dat after Ascension Day

Final project

Deadline for handing in: Wednesday, 31–5, 23.59.

Chapter 7

Testing

An overview of the testing is given in Table 7.1.

Grading

The final grade of this course will be composed of the grade of two mid-term take home exams (2 × 20%) and one individual project (60%).

Mid-term take-home exam $(2 \times 20\%$

In two mid-term take-home exam, students will show their understanding of the literature and prove they have gained new insights during the lecture/seminar meetings. They will be asked to critically assess various approaches to Big Data analysis and make own suggestions for research. Additionally, they need to (partly) write the code to accomplish this.

Grading criteria are communicated to the students together with the assignment, but in general are: For literature-related tasks in the exam:

- usage of specific examples from the literature;
- critique of different approaches;
- nameing of pro's, con's, potential pitfalls, and alternatives;
- giving practical advice and guidance.

For programming-related tasks in the exam:

- correctness, efficiency, and style of the code
- correctness, completeness, and usefulness of analyses applied.

Table 7.1: Test matrix

	In-class assignments, reviewing work of fellow students, active participation (precondition)	Mid-term take home exams (2×20% of final grade)	Final individual project (60% of final grade)
A. Students can explain the research designs and methods employed in existing research articles on Big Data and automated content analysis.	X	X	
B. Students can on their own and in own words critically discuss the pros and cons of research designs and methods employed in existing research articles on Big Data and automated content analysis; they can, based on this, give a critical evaluation of the methods and, where relevant, give advice to improve the study in question.	X	X	
C. Students can identify research methods from computer science and computer linguistics which can be used for research in the domain of communication science; they can explain the principles of these methods and describe the value of these methods for communication science research.4. Skills and abilities: Are able, independently and on their own, to set up, conduct, report and interpret advanced academic research.	X	X	X
D. Students can on their own formulate a research question and hypotheses for own empirical research in the domain of Big Data.			X
E. Students can on their own chose, execute and report on advanced research methods in the domain of Big Data and automatic content analysis.			X
F. Students know how to collect data with scrapers, crawlers and APIs; they know how to analyze these data and to this end, they have basic knowledge of the programming language Python and know how to use Python-modules for communication science research.	X	X	X
G. Students can critically discuss strong and weak points of their own research and suggest improvements.			X
H. Students participate actively: reading the literature carefully and on time, completing assignments carefully and on time, active participation in discussions, and giving feedback on the work of fellow students give evidence of this.	X		

For conceptual and planning-related tasks:

- feasibility
- level of specificity
- explanation and argumentation why a specific approach is chosen
- creativity.

Final individual project (60%)

The final individual project typically consists of the following elements, which all contribute to the final grade:

- introduction including references to relevant (course) literature, an overarching research question plus subquestions and/or hypotheses (1–2 pages);
- an overview of the analytic strategy, referring to relevant methods learned in this course;
- carefully collected and relevant dataset of non-trivial size;
- a set of scripts for collecting, preprocessing, and analyzing the data. The scripts should be well-documented and tailored to the specific needs of the own project;
- output files;
- a well-substantiated conclusion with an answer to the RQ and directions for future research.

Depending on the choosen topic, the student will have to apply multiple, but not all, techniques covered in the course. In particular, the student needs to spend a substantial amount of the project on a technique covered in Part II of this course. Student and teacher discuss the scope of the projects, the requirements that the specific project suggested by the student needs to fulfill, and the extend to which the different methods that the student plans to use will contribute to the final grade.

Grading and 2$^{\text{nd}}$ try

Students have to get a pass (5.5 or higher) for both mid-term take-home exams and the individual project. If the grade of one of these is lower, an improved version can be handed in within one week after the grade is communicated to the student. If the improved version still is graded lower than 5.5, the course cannot be completed. Improved versions of the final individual project cannot be graded higher than 6.0.

Chapter 8

Lecturers' team, including division of responsibilities

dr. Damian Trilling

Chapter 9

Calculation of students' study load (in hours)

- Elective total: 12 ECTS = 336 hours
- Reading:
  - 16 articles, average 20 pages: 320 pages. 6 pages per hour, thus 53 hours for the literature
  - Reading and doing tutorials: 80 hours for reading tutorials to acquire skills.
  - Reading book: 20 hours
  - Reading/preparation total: 153 hours.
- Presence:
  28*2 hours: 56 hours.
- Mid-term take-home exam, including preparation (2 exams) 2 × 14 hours: 28 hours
- Final individual project, including data collection, analysis, write up: 90 hours

Total: 337 hours

Chapter 10

Calculation of lecturers' teaching load (in hours)

- Presence: 56 hours (= 28 * 2 hours)
- Preparation of weekly lectures, 14 * 4 hours: 56 hours
- Preparation of weekly tutorials, 14 * 4 hours: 56 hours
- Assisting students with setting up Virtual Machine, individual help: 20 hours
- Feedback and grading take-home exams: 25x20 minutes x 2 exams: 17 hours
- Feedback and grading final projects, including feedback on proposal and individual counseling: 25* 60 min: 25 hours
- Administration, e-mails, individual appointments: 10 hours

Total: 240 hours

Chapter 11

History of the course

In response to the feedback by the test review committee in 2019, the following changes were applied:

- Empasized in the section “Exit qualification” that – in contrast to the 6 ECTS course – knowledge of *both* techniques from Part I (basic text processing, data retrieval from web sources) and Part II (supervised machine laring, and unsupervised machine learning) need to be demonstrated in the final project.
- Empasized the same in the section “Grading”
- Added specific grading criteria to the section “Grading”

In 2022, the course received a thorough update due to the publication of the new textbook (van Atteveldt et al., 2022).

In 2023, the course was updated to incorporate recent developments in machine learning. This was achieved by devoting less attention to now obsolete techniques that are mainly of historical interest, but are not considered best practice any more.

Literature

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* (Vol. 1) (No. 1). Association for Computing Machinery. doi: 10.1145/3442188.3445922

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant autmated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. doi: 10.1080/21670811.2015.1096598

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data. *Information, Communication & Society*, *15*(5), 662-679. doi: 10.1080/1369118X.2012.678878

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word Embeddings as a Lens to Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. doi: 10.1073/pnas.1720347115

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hilbert, M., Barnett, G., Blumenstock, J., Contractor, N., Diesner, J., Frey, S., ... Zhu, J. J. H. (2019). Computational Communication Science : A Methodological Catalyzer for a Maturing Discipline. *International Journal of Communication*, *13*, 3912–3934.

Hovy, D. (2020). *Text analysis in Python for social scientists: Discovery and exploration.* Cambridge, UK: Cambridge University Press. doi: 10.1017/9781108873352

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. doi: 10.1177/2053951714528481

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). From Word Embeddings To Document Distances. *Proceedings of The 32nd International Conference on Machine Learning*, *37*, 957–966.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., ... van Alstyne, M. (2009). Computational social science. *Science*, *323*, 721–723. doi: 10.1126/science.1167742

Lin, Z., Welbers, K., Vermeer, S., & Trilling, D. (2023). Beyond discrete genres: Mapping news items onto a multidimensional framework of genre cues. In *International Conference on the Web and Social Media (ICWSM).* (`https://arxiv.org/abs/2212.04185`)

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

Nothman, J., Qin, H., & Yurchak, R. (2018). Stop Word Lists in Free Open-source Software Packages. In *Proceedings of workshop for nlp open source software (nlp-oss)* (pp. 7–12). Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.18653/v1/W18-2502

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Salganik, M. J. (2017). *Bit by bit: Social research in the digital age.* Princeton, NJ: Princeton University Press.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *9th Python in science conference.*

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big Data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13. doi: 10.1177/0002716215572084

van Atteveldt, W., Trilling, D., & Arcila Calderón, 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.

van Atteveldt, W., & Peng, T. Q. (2018). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, *12*(2-3), 81–92. doi: 10.1080/19312458.2018.1458084

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data.* Sebastopol, CA: O'Reilly.