Big Data and Automated Content Analysis (6 ECTS)

Cursusdossier

dr. Anne Kroon

Graduate School of Communication
University of Amsterdam

a.c.kroon@uva.nl

Office: REC-C, 7th floor

Academic Year 2018/19
Semester 2, block 2

Contents

Chapter 1

Short description of the course

"Big data" is a relatively new phenomenon, and 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 even *Comutational 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.

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. Some basic understanding of programming languages is helpful, but not necessary to enter the course. Students without such knowledge are encouraged to follow the (free) online course at `https://www.codecademy.com/learn/python` to prepare.

Chapter 2

Exit qualifications

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 sixteen 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 two summative graded assignments: a 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 Trilling (2019) will be dealt with. Note that some basic chapters, which provide the students with the computer skills necessary to use our tools and explain which software to install, 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 the first three books use slightly outdated examples (e.g., Python 2, now-defunct APIs etc.).

- Russel, 2013. Gives a lot of examples about how to analyze a variety of online data, including Facebook and Twitter, but going much beyond that.

- Bird, Loper, & Klein, 2009. This is the official documentation of the NLTK package that we are using. A newer version of the book can be read for free at `http://nltk.org`

- McKinney, 2012: Another book with a lot of examples. A PDF of the book can be downloaded for free on `http://it-ebooks.info/book/1041/`.

- VanderPlas, 2016: A more recent 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`

- 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: PREPARING YOUR COMPUTER
Follow all steps as outlined in Chapter 1.

Week 1: Introduction

Monday, 1-4. Lecture.

(Anne)

We discuss what Big Data means, how the concept can be understood, what challenges and opportunities arise, and what the implications are for communication science.

Mandatory readings (in advance): boyd & Crawford, 2012, Kitchin, 2014.

Additional literature, not obligatory to read in advance, but very informative: Mahrt & Scharkow, 2013, Vis, 2013, Trilling, 2017.

Thursday, 4–4. Lab session.

✔ CHAPTER 2: THE LINUX COMMAND LINE
✔ CHAPTER 3: A LANGUAGE, NOT A PROGRAM
(Anne)
We will get familiar with the Virtual Machine and the software we will work with. Make sure you installed everything in advance and that you can start up your machine.

Week 2: Getting started with Python

Monday, 8–4. Lecture.

✔ Chapter 4: The very, very basics of programming in Python
(Anne)
You will get a very gentle introduction to computer programming. During the lecture, you are encouraged to follow the examples on your own laptop.

Thursday, 11–4. Lab session.

✔ Appendix A: Exercise 1
(Anne)
We will do our first real steps in Python and do some exercises to get the feeling.

Week 3: Data harvesting and storage

This week is about data sources and their (dis)advantages.

Monday, 15–4. Lecture.

(Anne)
A conceptual overview of APIs, scrapers, crawlers, RSS-feeds, databases, and different file formats.

Read the article by Morstatter, Pfeffer, Liu, and Carley (2013) in advance. It discusses the quality of data provided by the Twitter API. As a practical example for how "dirty" input data (i.e., data that for whatever reason does not come in form of a clean, structured data set like a table) can be parsed and preprocessed, have a look at the method section of the article by Lewis, Zamith, and Hermida (2013).

Thursday, 18–4. Lab session.

✔ Chapter 5.1–5.3: Retrieving and storing data
(Anne)
We will write a script to collect some data.

Week 4: Sentiment analysis

Up till now, we have mainly talked about available data and how to acquire them. From now on, we will focus on analyzing them and cover one technique per week. By now, you should also have gotten some idea about your final project.

Monday, 22–4. Holiday (Easter)

Thursday, 25–4. Lecture.

(Anne)
We start with an overview of different analytical approaches which we will cover in the next weeks, After that, we will focus on the first of these techniques, sentiment analysis.

Read the following two articles in advance. The first one gives an overview of how to analyze social media data, in this case, Twitter (Bruns & Stieglitz, 2013). The other one is an example of a sentiment analysis (Mostafa, 2013).

Some additional examples of sentiment analysis (not obligatory): Huang, Goh, and Liew (2007); Pestian et al. (2012). If you want to have a look under the hood of a popular sentiment analysis algorithm, you can read Thelwall, Buckley, and Paltoglou (2012) and Hutto and Gilbert (2014).

Week 5: Automated content analysis with NLP and regular expressions

Text as written by humans usually is pretty messy. You will learn how to process text to make it suitable for further analysis by using techniques of Natural Language Processing (NLP), and how to extract meaningful information (discarding the rest) using regular expressions.

Monday, 29–4. Lab session (accompanying Lecture week 4).

✔ Chapter 6: Sentiment analysis
(Anne)
You will write a tool to read data and conduct a sentiment analysis.

Thursday, 2–5. Lecture with exercises

✔ Chapter 7: Automated content analysis
(Anne)
This lecture will introduce you to techniques and concepts like stemming, stopword removal, n-grams, word counts and word co-occurrances, and regular expressions. We will do some exercises during the lecture.

Preparation: Mandatory reading: Boumans & Trilling, 2016. Also read the paper by Madnani (n.d.). It uses the same package (NLTK) which we use in class, If you don't get all practical details yet, that's OK. Pay special attention to the (linguistic) concepts applied.

Take-home exam

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

Week 6: Web scraping and parsing

Monday, 6–5. Lecture.

(Anne)
We will 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.

Thursday, 9–5. Lab session

✔ Chapter 8: Web scraping
(Anne)
We will exercise with web scraping and parsing.

Week 7: Statistics, Machine Learning, and what to do next

Monday, 13–5. Short lecture plus lab session.

✔ Section 3.5: Jupyter Notebook
✔ Chapter 12: Statistics with Python
(Anne)
You have worked hard so far, so we'll do something fun and relaxing (of course, fun might be a relative concept in this course. . .). You are going to learn how to create visualizations, do conventional statistical tests, manage datasets with Python, save the results together with your code and your own explanations – and all of this within your browser.

Thursday, 16–5. Lecture

(Anne)
This lecture will introduce you to one of the most fascinating topics in automated content analysis: machine learning. I will walk you trough the ideas behind unsupervised and supervised machine learning. The nice thing is that you actually have already done it during your studies: Principal component analysis is a form of unsupervised ML and regression analysis a form of supervised ML – you just never called it like this. And you probably never thought about using these techniques to analyze texts (or images). And that's what we are going to do. We will not be able to do all of this in this course, but you learn where to look for resources in case you want to go on with computational methods.

Finish!

This week is designated to working on your final projects.

Monday, 20–5. Open Lab (not mandatory)

(Anne)
Possibility to ask last questions regarding the final project.

Final project

Deadline for handing in: Wednesday, 29–5, 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 one mid-term take home exam (30%) and one individual project (70%).

Mid-term take-home exam (30%)

In a 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.

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 exam (30% of final grade)	Final individual project (70% 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 (70%)

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. 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 2nd try

Students have to get a pass (5.5 or higher) for both the mid-term take-home exam 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 (responsible)
dr. Theo Araujo assists during practicum sessions (30 extra hours, not included in calculation below; necessary because it is unfeasible to help > 15 students individually during practicum sessions)

Chapter 9

Calculation of students' study load (in hours)

- Elective total: 6 ECTS =168 hours
- Reading:
  - 8 articles, average 20 pages: 160 pages. 6 pages per hour, thus 26 hours for the literature
  - Reading and doing tutorials: 14 hours for reading tutorials to acquire skills.
  - Reading book: 20 hours
  - Reading/preparation total: 60 hours.
- Presence:
  14*2 hours: 28 hours.
- Mid-term take-home exam, including preparation: 14 hours
- Final individual project, including data collection, analysis, write up: 66 hours

Total: 168 hours

Chapter 10

Calculation of lecturers' teaching load (in hours)

- Presence: 28 hours (= 14 * 2 hours)
- Preparation of weekly lectures: 7 * 4 hours: 28 hours
- Preparation of weekly lab meetings: 7 * 4 hours: 28 hours
- Assisting students with setting up Virtual Machine, individual help: 20 hours
- Feedback and grading take-home exams: 25 * 20 minutes: 8 hours
- Feedback and grading final projects, including feedback on proposal and individual counseling: 25 * 60 min: 25 hours
- Administration, e-mails, individual appointments: 9 hours

Total: 146 hours

Literature

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly.

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

Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: metrics for tweeting activities. *International Journal of Social Research Methodology*, *16*(2), 91–108. doi: 10.1080/13645579.2012.756095

Huang, Y.-P., Goh, T., & Liew, C. L. (2007). Hunting suicide notes in Web 2.0 – preliminary findings. *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*, 517–521. doi: 10.1109/ISM.Workshops.2007.92

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international aaai conference on weblogs and social media.*

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

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

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of Big Data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, *57*(1), 34–52. doi: 10.1080/08838151.2012.761702

Madnani, N. (n.d.). *Getting started on natural language processing with Python.* http://desilinguist.org/pdf/crossroads.pdf.

Mahrt, M., & Scharkow, M. (2013). The value of Big Data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20–33. doi: 10.1080/08838151.2012.761700

McKinney, W. (2012). *Python for data analysis.* Sebastopol, CA: O'Reilly.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from Twitter's Streaming API with Twitter's Firehose. In *International AAAI conference on weblogs and social media (ICWSM).* Boston, MA. Retrieved from `http://www.public.asu.edu/~fmorstat/paperpdfs/icwsm2013.pdf`

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, *40*(10), 4241–4251. doi: 10.1016/j.eswa.2013.01.019

Pestian, J., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., … Brew, C. (2012). Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, *5*, 3–16. doi: 10.4137/BII.S9042

Russel, M. (2013). *Mining the social web. Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more* (2nd ed.). Sebastopol, CA: O'Reilly.

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

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

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173. doi: 10.1002/asi.21662

Trilling, D. (2017). Big Data, Analysis of. In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *The international encyclopedia of communication research methods.* Wiley. doi: 10.1002/9781118901731.iecrm0014

Trilling, D. (2019). Doing computational social science with Python: An introduction. Version 1.3. *SSRN*. Retrieved from `http://papers.ssrn.com/abstract=2737682`

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

Vis, F. (2013). A critical reflection on Big Data: Considering APIs, researchers and tools as data makers. *First Monday*, *18*(10), 1–16. doi: 10.5210/fm.v18i10.4878