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

Course Manual

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 2018/19
Semester 2, block 1 and 2

Chapter 1

About this course

This course manual contains general information, guidelines, rules and schedules for the Research Master course Big Data & Automated Content Analysis Part I and II (12 ECTS). Please make sure you read it carefully, as it contains information regarding assignments, deadlines and grading.

1.1 Course description

"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. 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 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.

1.2 Goals

Upon completion of this course, the following goals are reached:

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 computational 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.

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.

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.

1.3 Help with practical matters

While making your first steps with programming in Python, you will probably have a lot of questions. Nevertheless, `htttp://google.com` and `http://stackoverflow.com` should be your first points of contact. After all, that's how we solve our problems as well...

Chapter 2

Rules, assignments, and grading

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

2.1 Mid-term take-home exams ($2 \times 20\%$)

In two mid-term take-home exam, students will show their understanding of the literature and prove they have gained new insights during the lectures and lab sessions. They will be asked to critically assess various approaches to Big Data analysis and make own suggestions for research.

2.2 Final individual project (60%)

The final individual project typically consists of the following elements:

- 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.

2.3 Grading and 2nd 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.

2.4 Presence and participation

Attendance is compulsory. Missing more than three meetings – for whatever reason – means the course cannot be completed.

Next to attending the meetings, students are also required to prepare the assigned literature and to continue working on the programming tasks after the lab sessions. To successfully finish the course, attending the lab sessions is not enough, but has to go hand-in-hand with continuos self-study.

2.5 Staying informed

It is your responsibility to check the means of communications used for this course (i.e., your email account, but – if applicable – also e-learning platforms or any other tool that the lecturer decides to use) on a regular basis, which in most cases means daily.

2.6 Plagiarism & fraud

Plagiarism is a serious academic violation. Cases in which students use material such as online sources or any other sources in their written work and present this material as their own original work without citation/referencing, and thus conduct plagiarism, will be reported to the Examencommissie of the Department of Communication without any further negotiation. If the committee comes to the conclusion that a student has indeed committed plagiarism the course cannot be completed.

General UvA regulations about fraud and plagiarism apply.

2.7 Deadlines and handing in

Please send all assignments and papers as a PDF file to ensure that it can be read and is displayed the same way on any device. Hardcopies are not required. Multiple files should be compressed and handed in as one .zip file or .tar.gz file. Anything exceeding a reasonable file size (approximately 5 MB) has to be send via `https://filesender.surf.nl/` instead of direct email.

Final papers and take-home exams that are not handed in on time, will be not be graded and receive the grade 1. This rule also applies for any other assignment that might be given. The deadline is only met when the all files are submitted.

Chapter 3

Schedule and 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.

Before the course starts: Prepare your computer.

✔ Chapter 1: Preparing your computer
Follow all steps as outlined in Chapter 1.

PART I: Basics of Python and ACA

Week 1: Introduction

Wednesday, 6–2. Lecture.

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 and Crawford (2012) and Kitchin (2014).

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

Friday, 8–2. No meeting.

✔ Chapter 2: The Linux command line
✔ Chapter 3: A language, not a program

Read the two chapters, and make sure you can reproduce the examples on your computer. Write down specific questions you have, so that you can ask them on Monday. It is encouraged to do so in pairs or groups.

Week 2: Getting started with Python

Wednesday, 13–2. Lecture.

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

Friday, 15–2. Lab session.

✔ Appendix A: Exercise 1
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.

Wedneday, 20–2. Lecture.

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).

Friday, 22–2. Lab session.

✔ Chapter 5.1–5.4: Retrieving and storing data
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.

Wednesday, 27–2. Lecture.

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.

Mandatory readings (in advance): Gonzalez-Bailon and Paltoglou (2015) and Hutto and Gilbert (2014).

Suggestions for additional readings:

- Examples of (simple) sentiment analyses: Huang, Goh, and Liew (2007); Mostafa (2013); Pestian et al. (2012).
- If you want to have a look under the hood of another popular sentiment analysis algorithm, you can read Thelwall, Buckley, and Paltoglou (2012).

Friday, 1–3. Lab session.

✔ Chapter 6: Sentiment analysis
You will write a script to read data and conduct a sentiment analysis.

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. You will also make a first aquintance with the packages NLTK and spacy.

Wedneday, 6–3. Lecture with exercises.

✔ Chapter 7: Automated content analysis
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 and Trilling (2016).

Friday, 8–3. Lab session.

You will combine the techiques discussed on Wednesday and write a full automated content analysis script using a top-down dictioary or regular-expression approach.

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 (12–5, 23.59).

Week 6: Web scraping and parsing

Wednesday, 13–3. Lecture.

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.

Friday, 15–3. Lab session.

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

Week 7: Statistics with Python

Wednesday, 20–3. Short lecture plus lab session.

✔ Section 3.5: Jupyter Notebook
✔ Chapter 12: Statistics with Python
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.

Friday, 22–3. Short lecture plus lab session.

We will learn how to do data wrangling with pandas: converting between wide and long formats (melting and pivoting), aggregating data, joining datasets, and so on.

– Break between block 1 and 2 –

PART II: Advanced analyses

Week 8: Dealing with temporal data

Wednesday, 3–4. Guest lecture by Rens Vliegenthart.

We will talk about time series analysis. Many data you can collect online have some temporal component in them: think of references to political parties or topics over time, or of the coverage of an organization. The same holds true for non-media data, such as stock exchange rates or unemployment statistics. We will discuss statistical models to analyse such data.

Mandatory reading (in advance): Vliegenthart (2014) and Strycharz, Strauss, and Trilling (2018).

Friday, 5–4. Lab session.

We will use an example to explore how to implement the techniques discussed on Wednesday using statsmodels (Seabold & Perktold, 2010).

Week 9: Supervised Machine Learning 1

In weeks 9 and 10, you will learn how to work with scikit-learn (Pedregosa et al., 2011), one of the most well-known machine learning libraries.

Wednesday, 10–4. Lecture.

We will learn the principles of supervised machine learning and discuss how logistic regression and Naive Bayes classifiers can be used to predict, for instance, movie ratings or topics of news articles. We will also discuss basics evaluation metrics like precision and recall.

Mandatory reading (in advance): Burscher, Odijk, Vliegenthart, de Rijke, and de Vreese (2014).

Friday, 12–4. Lab session.

✔ Chapter 10: Supervised machine learning
You will build your first machine learning classifier.

Week 10: Supervised Machine Learning 2

Wednesday, 17–4. Lecture with practical exercise.

We will discuss more in detail how to select the best model for your purpose. We will talk about cross-validation, parameter tuning, and building a pipeline.

Friday, 19–4. No meeting (Easter)

Week 11: Unsupervised Machine Learning 1

Wednesday, 24–4. Lecture.

We will discuss the basics of unsupervised machine learning, using techniques such as principal component analysis, k-means clustering, and hiearchical clustering.

Mandatory reading (in advance): Burscher, Vliegenthart, and de Vreese (2016).

Friday, 26–4. Lab session.

You will apply the techniques discussed on Wednesday.

Take-home exam

In week 11, 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 (30–4, 23.59).

Week 12: Unsupervised Machine Learning 2

Wednesday, 1–5

We will discuss one of the most popular unsupervised techniques in automated content analysis: topic modeling. In particular, we will focus on LDA.

Mandatory readings (in advance): Maier et al. (2018) and Tsur, Calacci, and Lazer (2015).

Friday, 3–5

✔ Chapter 11: Unsupervised machine learning
You will apply the techniques discussed on Wednesday using gensim (Řehůřek & Sojka, 2010).

Week 13: Word embeddings

Wednesday, 8–5

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.

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

Friday, 10–5

We will apply a word2vec model.

Week 14: Wrapping up & moving on

Wednesday, 15–5. Lecture

In this meeting, we will wrap up what has been covered in this course and discuss what other techniques and approaches exist that we did not have time to cover in detail, such as deep learning.

Friday, 17–5. Open Lab

Possibility to ask last questions regarding the final project.

Final project

Deadline for handing in: Wednesday, 29–5, 23.59.

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

Burscher, B., Odijk, D., Vliegenthart, R., de Rijke, M., & de Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, *8*(3), 190–206. doi: 10.1080/19312458.2014.937527

Burscher, B., Vliegenthart, R., & de Vreese, C. H. (2016). Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review*, *34*(5), 530-545. doi: 10.1177/0894439315596385

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

Gonzalez-Bailon, S., & Paltoglou, G. (2015). Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 95–107. Retrieved from `http://ann.sagepub.com/content/659/1/95.abstract?rss=1` doi: 10.1177/0002716215569192

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

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

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

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

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

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

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.

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

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. (`http://is.muni.cz/publication/884893/en`)

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.

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

Strycharz, J., Strauss, N., & Trilling, D. (2018). The role of media coverage in explaining stock market fluctuations: Insights for strategic financial communication. *International Journal of Strategic Communication*, *12*(1), 67–85. doi: 10.1080/1553118X.2017.1378220

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`

Tsur, O., Calacci, D., & Lazer, D. (2015). A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (pp. 1629–1638). ACL.

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

Vliegenthart, R. (2014). Moving up. Applying aggregate level time series analysis in the study of media coverage. *Quality & Quantity*, *48*(5), 2427–2445. doi: 10.1007/s11135-013-9899-0