Four-day workshop Automated Content Analysis with Python

Course Manual

dr. Damian Trilling

Department of Communication Science
University of Amsterdam

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

11/12 & 25/26 September 2017

About this course

Course description

In the social sciences, there is an increasing interest for automatically analyzing texts. This in particular concerns research that draws on Internet-based data sources such as social media, online news, 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 *Computational Communication Science* (Shah, Cappella, & Neuman, 2015).

The workshop will provide insights in the basic concepts, challenges and opportunities associated with data so large that traditional research methods (like manual coding) cannot be applied any more. Participants are introduced to strategies and techniques for analyzing large quantities of text, through concrete examples and templates than can be shared and modified for own research projects.

Goals

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

A Participants can identify research methods from computer science and computational linguistics which can be used for research in the domain of the social sciences; they can explain the principles of these methods and apply them to the analysis of texts.

B Participants have a basic knowledge of the Python programming language and can work with Jupyter Notebooks.

C Participants can at least on a basic level implement techniques from (A), using commonly used Python modules.

Readings

We will work with the book by Trilling (2017), which can be downloaded as a PDF file (see references).

In the schedule below, I refer to chapters from this book, where you can find a bit more background and explanation about the topics we will discuss..

Preparation and Prerequisits

✔ Chapter 1: Preparing your computer ***or (!)*** install Anaconda

You are expected to bring a laptop with either Anaconda (Python 3 version) *or* a virtual machine as explained in Chapter 1 of the book installed. Anaconda can be downloaded from `https://www.continuum.io/downloads`.

Monday, 11-7-2017

Morning (9.30–12.30)

✔ Chapter 3: A language, not a program, in particular:
✔ Section 3.5: Jupyter Notebook
✔ Section 4.1: Data types

After a short introduction round, we will discuss some basic characteristics of Python, such as:

- Why Python?
- How does Python relate to software you might be familiar with, such as SPSS, STATA, or R?
- An introduction to data types – or why a "variable" is not what you might think

Finally, we'll play around a bit with different Python interpreters.

Afternoon (13.30–16.30)

✔ Chapter 4: The very, very basics of programming in Python

In the afternoon, you will learn about the basics of programming. We will talk about functions, methods, loops, and much more.

Tuesday, 12-7-2017

Morning (9.30–12.30)

✔ Boumans and Trilling (2016)
✔ Gonzalez-Bailon and Paltoglou (2015)

✔ Chapter 6: Sentiment analysis

We will start with an overview about different approaches to automated content analysis (ACA), as outlined by Boumans and Trilling (2016).

We will discuss sentiment analysis, as it is one of the frequently applied out-of-the-box techniques for analyzing text. We will discuss the simple word count methods, off-the-shelf modules such as Vader (Hutto & Gilbert, 2014) and Pattern (De Smedt & Daelemans, 2012), and briefly talk about more sophisticated alternatives.

Afternoon (13.30–16.30)

✔ Chapter 7: Automated content analysis

After that, we will move on to techniques that can be better tailored to your own research questions. Text as written by humans usually is pretty messy. You will therefore 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. Also, I will introduce you to techniques and concepts like stemming, stopword removal, n-grams, word counts and word co-occurrances.

Monday, 25-5-2017

Morning (9.30–12.30)

✔ Chapter 10: Supervised machine learning

In the first meeting, we considered techniques that are rule-based and, in general, deterministic: “if you encounter X, do Y”, “if you find string X, code as mention of actor A”. While such techniques can be a informative and are necessary to automatically and efficiently perform routine tasks, they are less suitable for offering really deep insights into what texts are about.

In this session, I will therefore introduce you to one of the most fascinating topics in automated content analysis: Machine learning.

I will walk you trough the ideas behind supervised machine learning. The nice thing is that you actually have already done it before: regression analysis a form of supervised ML – you just never called it like this. And you probably never thought about using such a technique to analyze texts (or images). And that’s what we are going to do, using one of the most commonly used frameworks for this, the Python package *scikit-learn* (Pedregosa et al., 2011).

Afternoon (13.30–16.30)

✔ Chapter 11: Unsupervised machine learning

In the morning session, we learned a technique for analyzing data assuming that we have a labeled training dataset. But what if we don't? That's where unsupervised machine learning comes into play. Also this is a principle you know already: Principal component analysis is a form of unsupervised ML. We will use a more sophisticated model, though, Latent Dirichlet Allocation (LDA) – which is a form of so-called topic modelling. We will use the Python package *gensim* (Řehůřek & Sojka, 2010) as well as the visualization module *pyldavis* (Sievert & Shirley, 2014).

Tuesday, 26-5-2017

Morning (9.30–12.30)

Afternoon (13.30–16.30)

✔ Chapter 8: Web scraping

We will conclude the workshop with a whole day of what might be the most challenging part: getting actual data! We will look into how to scrape, for instance, review sites and similar online services.

Further readings

There are a lot of online ressources available for using Python in the social sciences, just google it.

Some jupyter notebooks that might be interestig for you are available at `http://damiantrilling.net/tools`.

Next to this, the following books provide the interested participants with more and deeper information. They are intended for the advanced reader and might be very useful for those who want to go deeper into the topic. Keep in mind that, due to the rapidly changing nature of the subject, parts of them are outdated already.

- 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/`.

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

De Smedt, T., & Daelemans, W. (2012). Pattern for Python. *The Journal of Machine Learning Research*, *13*, 2063–2067.

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. doi: 10.1177/0002716215569192

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.*

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

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

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.

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.

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

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. Retrieved from `http://www.aclweb.org/anthology/W/W14/W14-3110` doi: 10.1.1.100.1089

Trilling, D. (2017). Doing computational social science with Python: An introduction. Version 1.0. *SSRN*. Retrieved from `http://papers.ssrn.com/abstract=2737682`