Two-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

Office: REC-C, 8th floor

31-1-2016 & 2-2-2016

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, we will refer to chapters from this book.

Preparation and Prerequisits

✔ Chapter 1: Preparing your computer ***or (!)*** install Anaconda
✔ Section 3.5: Jupyter Notebook

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

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;
- have a (very) basic understanding of the Python programming language;
- know how to work with Jupyter Notebook (on your computer).

Meeting 1

✔ Chapter 6: Sentiment analysis
✔ Chapter 7: Automated content analysis

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

We will briefly discuss sentiment analysis, as it is one of the frequently applied out-of-the-box techniques for analyzing text.

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.

Meeting 2

✔ Chapter 10: Supervised machine learning
✔ Chapter 11: Unsupervised 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 unsupervised and supervised machine learning. The nice thing is that you actually have already done it before: 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.

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`.

Suggested additional literature on sentiment analysis:

- Some examples of sentiment analysis: Huang, Goh, and Liew (2007); Mostafa (2013); Pestian et al. (2012).
- If you want to have a look under the hood of popular sentiment analysis algorithms, you can read Thelwall, Buckley, and Paltoglou (2012) and Hutto and Gilbert (2014).

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

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.*

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.

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.

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