Workshop Automated Content Analysis

Johannes B. Gruber

2023-07-06

Overview

We are going to look at the different topics mostly from a practical standpoint with a little theoretical and statistical background where necessary. The course will deal with the following topics:

time	Day 1	Day 2
09:00-10:30	Obtaining Text Data	Word Embeddings
11:00:12:30	Text Scaling and Regression Models	Deep Learning
14:00-15:30	Supervised Classification Methods	Big Data Projects: Some Tips

Introduction

The availability of text data has exploded in the last two decades. First the availability of text through digital archives, then the advent of digital media communication like online news and press releases and most recently public communication of non-elite actors on social media. For political science this opens up exciting new possibilities for research as many processes which occurred in private or elite venues is now accessible. At the same time, the sheer amount of data makes manually analysing meaningul fractions of it impossible.

This course is an introduction to the available methods and software for automated content analysis. The 101 in it's name is meant to indicate that this is a introductory course. However, the introductory part is into automated content analysis while the expectation is that you are comfortable with R, the programming language used in this course.

What should be clear about the course from the beginning though is that despite recent advances, "All Quantitative Models of Language Are Wrong—But Some Are Useful" (Grimmer and Stewart 2013, 3). The primary goal of this course is thus to understand the types of questions we can ask with text, and how to go about answering them.

Overview, Background and some Theory

This session focuses on the general concepts in ACA, like pre-processing, the documents-term-matrix, dimensionality reduction and so on. It also provides a general overview on ACA-methods, how they are implemented in software and what kinds of research questions and designs are possible (or at least which have been asked before).

Readings:

1. Taking Stock of the Toolkit (Boumans and Trilling 2016)
2. Text Analysis in R (Welbers, van Atteveldt, and Benoit 2017)
3. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts (Grimmer and Stewart 2013)
4. Computer-Assisted Text Analysis for Comparative Politics (Lucas et al. 2015)

Obtaining Text Data

There are a myriad of ways to analyse text in `R`. If you ever want to make use of them though you have to somehow get your own data into `R`. This can be a bit boring and so this session might not be the most impressive one. But by the end of it, you will be able to use your own data in the coming sessions. And isn't that exciting!

Key Points:

- Reading in common file formats (txt, PDF, docx and so on).
- Case 1: Use of newspaper data
- Case 2: Web-Scraping (a brief overview)
- Case 3: Twitter scraping (How to make an account, install rtweet)

Readings:

- none; but think about what sources of text data you want to use and bring it along if possible.

Dictionary Methods

Dictionary approaches belong to the oldest and simplest methods in ACA. The key concept here is the dictionary, which is a list of words along with a category, such as positive/negative sentiment, anger/joy, geo-locations and so on. By checking if texts contain words from the category, we can infer if each text belongs to the category defined in the dictionary. In this session, we use a simple example and discuss the pro and cons of dictionary methods.

Additional Readings:

- Text Mining with R chapter 2 (Silge and Robinson 2020)

Text Scaling and Regression Models

One of the fundamental ideas of ACA is that text is just another form of data. Once we obtain text and turn it into a document-term-matrix, it is not fundamentally different from other forms of statistical data any more. Therefore we can perform all sorts of statistical analysis on it – like regressions. In political science, this fact inspired a technique called ideological scaling – one of the few methods discussed here that did not originate in statistics or computer science. The idea is to project texts (and by proxy the respective authors) onto a one- or two-dimensional space, often interpreted as a left-right political spectrum.

Additional Readings:

1. Supervised Machine Learning for Text Analysis in R 6 (Hvitfeldt and Silge 2021)
2. A Scaling Model for Estimating Time Series Party Positions from Texts (Slapin and Proksch 2008)

Supervised Classification Methods

The idea behind supervised classification or supervised learning approaches is that you train a model to emulate the behaviour of a human coder. Specifically, a human classifies texts into categories, such as positive/negative tone, spam/important emails and so on. By analysing the statistical distribution of words in the two or more categories, a model can predict the class of new unclassified material.

Readings:

- Supervised Machine Learning for Text Analysis in R 7 (Hvitfeldt and Silge 2021)

Unsupervised Classification Methods

Unsupervised classification or unsupervised learning is a type of machine learning where the computer is not given any labels or categories to assign to data. Instead, the computer is tasked with finding patterns and relationships in the data and then assigning categories to the data based on those patterns. This is done through a process called dimension reduction, which is similar to techniques like Principal Component Analysis (PCA) or factor analysis. To use this method, the researcher needs to define the number of categories they want the

computer to find and then interpret the results afterwards. One of the most popular methods for unsupervised classification is Latent Dirichlet Allocation (LDA) topic modeling. This method assigns a probability to each word in a corpus to belong to a certain topic, and then calculates the probability of each text in the corpus belonging to a certain topic. In this way, the computer can find patterns and relationships in the data and assign categories based on those patterns.

Additional Readings:

- Probabilistic topic models. (Blei 2012)
- Islamophobia and Media Portrayals of Muslim Women (Terman 2017)

Regular Expressions, String Hacking, Part-of-Speech Tagging

Working with text data often comes back to searching for or replacing certain patterns. Often these tasks can be accomplished by using a 'language' called regular expressions (shortened as regex). The idea is that special symbols or character strings can be used to make the computer find a number of different strings. The expression `"\bcat.*"` for example would match the strings "cat", "cats", "catastrophe" and many others. Regex patterns can be an incredibly powerful tool for processing your text. For example, you can remove or group similar words together if they share a common feature before doing any other analysis.

Part-of-Speech (POS) Tagging can solve a similar purpose (although it has many other uses we do not discuss here). POS finds grammatical features of words in a text, like if word is a noun, verb, adjective, adverb, etc.. Many POS-taggers can also determine if a noun is, for example, a person, a place or an entity. This can be useful if you want to restrict your analysis to certain words or remove certain entities that might obfuscate your findings.

Readings:

- R for Data Science chapter 14.3

Additional Readings:

- Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition chapter 2

Word Embeddings

This session introduces newer advances of text analysis that go beyond traditional bag-of-words models. Word embeddings are a way to represent words as vectors that capture their semantic meaning, and deep learning models use neural networks to process and analyze text data. Students will learn about popular word embedding algorithms like Word2Vec and GloVe,

as well as popular deep learning models for text analysis like CNNs and RNNs. Through demonstrations, students will learn how to use pre-trained word embeddings and implement simple deep learning models for text classification. The session will also explore real-world applications of these techniques in areas like sentiment analysis and text classification.

Addional Readings:

- Supervised Machine Learning for Text Analysis in R 8-10 (Hvitfeldt and Silge 2021)

Deep Learning

Deep learning is a subfield of machine learning that deals with algorithms inspired by the structure and function of the human brain, called artificial neural networks (ANNs). It has become especially prominent since the transformer deep learning architecture has been introduced by Vaswani et al. (2017). Since then large language models like BERT or GPT-3 (Generative Pre-trained Transformer 3) have outclassed previous approaches for text-as-data methods. This session will give an overview of some of the tools you need to use these models for your own analyses. These include:

- spaCy
- simpletransformers
- BERTopic

Big Data Projects: Some Tips

In the final session, we will focus on some general tips when running models that take longer than a few seconds to converge. These include:

- a good workflow with quarto documents
- "piloting" analysis steps
- running analysis on cloud infrsturcture

References

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84. https://doi.org/10.1145/2133806.2133826.

Boumans, Jelle W., and Damian Trilling. 2016. "Taking Stock of the Toolkit: An Overview of Relevant Automated Content Analysis Approaches and Techniques for Digital Journalism Scholars." *Digital Journalism* 4 (1): 8–23. https://doi.org/10.1080/21670811.2015.1096598.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Hvitfeldt, Emil, and Julia Silge. 2021. *Supervised Machine Learning for Text Analysis in R.* https://smltar.com/.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23 (2): 254–77. https://doi.org/10.1093/pan/mpu019.

Silge, Julia, and David Robinson. 2020. *Text Mining with R.* https://www.tidytextmining.com/.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–22. https://www.jstor.org/stable/25193842.

Terman, Rochelle. 2017. "Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage." *International Studies Quarterly* 61 (3): 489–502. https://doi.org/10.1093/isq/sqx051.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." arXiv. https://doi.org/10.48550/arXiv.1706.03762.

Welbers, Kasper, Wouter van Atteveldt, and Kenneth Benoit. 2017. "Text Analysis in R." *Communication Methods and Measures* 11 (4): 245–65. https://doi.org/10.1080/19312458.2017.1387238.