Text as data - syllabus

Max Callaghan

2022-09-08

Course contents and learning objectives

Course contents:

There is an abundance of unstructured data around us. Working with text is key, not just to measure discussions and opinions on social media or in product reviews, but also to gain insights into concepts important to the study of politics such as ideological positions or policy sentiments. The vast amount of textual data that one frequently needs to process, and the messy form that it often comes in, poses special challenges to researchers. This course introduces students to computational tools and methods that enable them to treat text as data. We will touch upon core theoretical concepts, but the main goal is to give you hands-on experience in using R to collect, load, prepare and analyze text data.

Main learning objectives

Students will gain hands-on experience working with different types of text data in R. This includes obtaining, managing, and wrangling data, as well as applying different models for analysing text

Target group

The course is aimed at students with an interest in programming, who wish to gain experience analysing large collections of texts.

Prerequisites

Basic knowledge of R is required

Diversity Statement

Understanding and respect for all cultures and ethnicities is central to the teaching at Hertie. Being mindful of diversity is an important issue for policy professionals in the planning, implementation, and evaluation of programmes designed for specific groups, populations, or communities. Diversity and cultural awareness will be integrated in the course content whenever possible.

Grading and Assignments

Assignment 1: Programming exercise	Deadline: 13 October	Submit via Moodle	30%
Assignment 2: data analysis exercise	Deadline:17 November	Submit via Moodle	30%
Assignment 3: Oral presentation of your own research project	Deadline 1 December	Presentation in class	40%

Assignment Details

Assignment 1

The first assignment will be a programming exercise, where you will receive a dataset and clear step-by-step instructions to import and manage text, and produce a simple analysis. Using what we have learnt in class, you will be required to write a script to carry out the instructions.

Assignment 2

The second assignment will involve the construction of a topic model, including a visualization and brief discussion of the results. Your grade for this assignment will be based on how you approach this task and if you follow the steps discussed in class.

Assignment 3

Your final assignment is the presentation of a group research project. You are free to choose your subject and methods, as long as the project involves the analysis of text and that the methods are covered in this course. Grading will be determined by the quality of the presentation, and the degree to which you manage to apply the skills you have learned during the course.

Late submission of assignments

For each day the assignment is turned in late, the grade will be reduced by 10 percentage points.

Attendance

Students are expected to be present and prepared for every class session. Active participation during lectures and seminar discussions is essential. If unavoidable circumstances arise which prevent attendance or preparation, the instructor should be advised by email with as much advance notice as possible. Please note that students cannot miss more than two out of 12 course sessions. For further information please consult the Examination Rules §10.

Academic Integrity

The Hertie School is committed to the standards of good academic and ethical conduct. Any violation of these standards shall be subject to disciplinary action. Plagiarism, deceitful actions as well as free-riding in group work are not tolerated. See Examination Rules §16 and the Hertie Plagiarism Policy.

Compensation for Disadvantages

If a student furnishes evidence that he or she is not able to take an examination as required in whole or in part due to disability or permanent illness, the Examination Committee may upon written request approve learning accommodation(s). In this respect, the submission of adequate certificates may be required. See Examination Rules §14.

Extenuating circumstances

An extension can be granted due to extenuating circumstances (i.e., for reasons like illness, personal loss or hardship, or caring duties). In such cases, please contact the course instructors and the Examination Office in advance of the deadline.

General Readings

Silge, Julia, and David Robinson. 2017. Text mining with R: A tidy approach. O'Reilly Media, Inc. https://www.tidytextmining.com/

Wickham, H. and G. Grolemund. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly. https://r4ds.had.co.nz/

Session Overview

Session 1: Text as data

The first session provides a general introduction to the subject, and outlines how we will proceed with the course.

Readings: Grimmer and Stewart (2013). 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents'. Political Analysis. link

Session 2: Preprocessing, and turning texts into features

This week, we will learn how we turn texts into features, and consider the choices we can make in doing so

Readings: Watanabe, Kohei and Stefan Müller. 2022. "Quanteda Tutorials". Chapter 3. https://tutorials.quanteda.io/basic-operations/.

Session 3: Acquiring, importing, and reading texts

In the third session we will cover how to acquire, and import text data, and we will demonstrate how to retrieve texts from APIs

Readings:

Bail, Chris. 2019. Text as Data: Application Programming Interfaces in R. https://cbail.github.io/textasdata/apis/rmarkdown/Application_Programming_interfaces.html

Session 4: Manipulating strings

This session will introduce you to regex expressions, and explain how the stringr package can be used to manipulate strings.

Readings:

Wickham, H. and G. Grolemund. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly. Chapter 14

Wickham, Hadley. 2010. "stringr: modern, consistent string processing". The R Journal 2 (2): 38-40.

Session 5: Visualising text features with ggplot

In this session, we will explore how to visualise the features we have generated from texts, and make graphs that show how corpora differ, or how they have changed over time.

Readings:

Silge, Julia, and David Robinson. 2017. Text mining with R: A tidy approach. O'Reilly Media, Inc. https://www.tidytextmining.com/. Chapter 3.

Assignment 1 set

Session 6: Word embeddings

In this session, we will learn about how word embeddings can offer richer representations of the content of texts

Readings:

Hvitfeldt, Emil, and Silge, Julia. 2019. Supervised Machine Learning for Text Analysis in R. Chapter 5 - Word Embeddings. https://smltar.com/embeddings.html

Assignment 1 due

Session 7: Dimensionality reduction

This session will demonstrate ways to represent multidimensional data in a 2-dimensional space.

Readings:

Conlen, Matthew, and Hohman, Fred. 2018. The Beginner's Guide to Dimensionality Reduction. https://dimensionality-reduction-293e465c2a3443e8941b016d.vercel.app/

Session 8: Topic modelling

This session will introduce topic models, and show how these can be run and visualised in R.

Readings:

Blei, David M. 2012. Probabilistic Topic Models. Communications of the ACM. 55 (4) pp 77–84 https://doi.org/10.1145/2133806.2133826 http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf

Assignment 2 set

Session 9: Sentiment analysis

This session will demonstrate how we can analyse the sentiment of texts using R.

Readings:

Silge, Julia, and David Robinson. 2017. Text mining with R: A tidy approach. O'Reilly Media, Inc. https://www.tidytextmining.com/. Chapter 2.

Session 10: Supervised learning

This session will demonstrate how we can use R to train machine learning models to reproduce a set of labels applied to texts

Readings:

Hvitfeldt, Emil, and Silge, Julia. 2022. Supervised Machine Learning for Text Analysis in R. CRC Press. https://smltar.com/. Chapter 7.

Assignment 2 due

Session 11: Spacy and Transformers

This session will explore some of the latest developments in NLP, and show some of the capabilities of Spacy and Transformers.

Readings:

Devlin, Jacob, et al. 2019. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of NAACL-HLT: 4171–4186. Alammar, Jay. The Illustrated Transformer: https://jalammar.github.io/illustrated-transformer/

Session 12: Wrapup and final presentations

In the final session, we will hear the presentations of the group projects you have conducted, and do a final wrap-up of what we have learned.

Assignment 3 due