Quantitative Text Analysis for Social Scientists

Tom Paskhalis

tom.paskhal.is

- **Module Code:** POP77032
- **Module Website:** tom.paskhal.is/POP77032
- **ECTS Weighting:** 10
- **Semester/Term Taught:** Semester 2 (Hillary Term)
- **Contact Hours:**
  One 2-hour lecture
  One 2-hour tutorial
  per week (11 weeks)
- **Module Coordinator:** Dr Tom Paskhalis (tom.paskhalis@tcd.ie)
- **Teaching Fellows:**
  - Sara Cid (cidsb@tcd.ie)

Learning Aims

At no time in human history has there been more textual information produced than the present day. Researchers now have access to massive collections of texts by different societal actors: parliamentary speeches and blog posts, corporate press releases and social media posts, newspaper articles and archival documents to name just a few. At the same time, the computational power has reached unprecedented levels and has enabled the development and use of practical software to process and analyze huge datasets of text.

This module focuses on a range of computational tools – stemming from the fields of machine learning and natural language processing (NLP) – that are essential for large-scale analyses of text information. The aim is to provide students with a hands-on introduction to processing and analyzing 'text-as-data' for the purpose of answering important social science research questions.

Learning Outcomes

On successful completion of this module students should be able to:

- understand the basic principles of treating text as data;
- extract and prepare textual data for analysis;
- apply key computational techniques for textual data;
- critically evaluate research that uses text analysis methods;

Prerequisites

This is an intermediate-level class focussing on representing text as quantitative data. The course assumes that you have a basic understanding of statistics and are comfortable with key programming concepts in R and/or Python.

Module Details

This module will consist of 2 parts: 2-hour lecture where we discuss approaches to empirical quantitative research and statistical methods, 2-hour tutorial where you have a chance to have hands-on experience working with data using `R` and `RStudio`.

In the course of this module students will submit 3 assignments that are designed to test their ability to (1) prepare the textual data for analysis and (2) apply the appropriate computational tools for extracting the key quantities of interest. The final assessment will be a short research paper where students will be asked to apply the techniques learned in the module to a research question of their choice.

Reading List

We will primarily be relying on the following core texts for this module:

- Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences.* Princeton, PA: Princeton University Press
- Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models.* 3rd. Draft. https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf

Some other useful texts on natural language processing and text analysis:

- Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* The MIT Press
- Jacob Eisenstein. 2019. *Introduction to Natural Language Processing.* Cambridge, MA: The MIT Press. https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

Finally, I highly recommend taking a look at the foundational content analysis text that was largely developed in pre-digital era but, nevertheless, provides an in-depth overview of many topics (largely pertaining to manual coding of text data) that are still highly relevant today:

- Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology.* 4th. Thousand Oaks, CA: SAGE Publications

In addition, we will use a number of journal articles. Most journal articles will be freely available from the link included in the reading list (from campus computers). If this does not work (or if you are not on campus), search for the article via Trinity Stella Search (or Google Scholar) and log in to gain access.

Additional online resources:

- quanteda
- APIs for Social Scientists: A Collaborative Review
- Text Mining with R

Software

In this class we will use R to work with data. R is free, open-source and interactive programming language for statical analysis. RStudio is a versatile editor for working with R code and data that provides a more intuitive interface to many features of the language.

Both R and RStudio are widely available for all major operating systems (Windows, Mac OS, Linux). You should install them on your personal computer prior to attending tutorials. Use these links to download the installation files:

- R - https://cran.r-project.org/
- RStudio - https://posit.co/download/rstudio-desktop/

Assessment Details

The final grade consists of the following parts (with corresponding weighting):

- Programming exercises (40% total)
- Research paper (60%)

  Approximately 3,000 words/5–10 pages (References and Appendix excluded)

In the research paper, each student will identify a research question and then answer it using computational text analysis tools. The data analysed in the paper should be textual in nature.

All assignments should be submitted via Blackboard. Go to the "Assessment" section — you should be able to see all the assignments listed there.

Please make sure that you understand the submission procedure. Unexcused late submissions will be penalized in accordance with standard department policy. Five points per day will be subtracted until the Monday a week and a half after the deadline at which point the assignment is deemed to have failed.