

POP77032 — Quantitative Text Analysis for Social Scientists

Module Information	
Module Code	POP77032
Module Title	Quantitative Text Analysis for Social Scientists
Department	Political Science
Year Group	Postgraduate (Taught Masters)
Module Personnel	Dr Tom Paskhalis
Academic Year	2026–27
Credit Weighting	10 credits (ECTS)
Semester	Semester 2 (Hilary Term)
Contact Hours	One 2-hour lecture and one 2-hour tutorial per week, over 11 teaching weeks.

Module Content	
Module Description	This module offers a hands-on introduction to the computational analysis of text data for the social sciences. Drawing on techniques from machine learning and natural language processing (NLP), it works through the full ‘text-as-data’ pipeline: acquiring and preparing textual data; representing text quantitatively; and applying dictionary-based, supervised and unsupervised methods, word embeddings, neural networks and large language models to answer substantive social science research questions. Each week combines a two-hour lecture with a two-hour hands-on tutorial in which students work with data using R and RStudio.
Module Learning Aims	To give students a practical, hands-on introduction to processing and analysing ‘text-as-data’, and to equip them with the computational tools needed to use large collections of text to answer 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;

Module Content	
	● extract and prepare textual data for analysis; ● apply key computational techniques for textual data; ● critically evaluate research that uses text analysis methods.
Indicative Reading / Resources	Core texts: ● Grimmer, J., Roberts, M. E. & Stewart, B. M. (2022) Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press. ● Jurafsky, D. & Martin, J. H. Speech and Language Processing (3rd edition). Further reading: ● Manning, C. D. & Schütze, H. (1999) Foundations of Statistical Natural Language Processing. MIT Press. ● Eisenstein, J. (2019) Introduction to Natural Language Processing. MIT Press. ● Krippendorff, K. (2019) Content Analysis: An Introduction to Its Methodology (4th edition). SAGE. Additional online resources include quanteda (quanteda.io), Text Mining with R (tidytextmining.com) and APIs for Social Scientists: A Collaborative Review. A full reading list is provided on the module website.

Assessment	
Assessment Components	● Programming Exercise 1 – 14% (coursework) ● Programming Exercise 2 – 13% (coursework) ● Programming Exercise 3 – 13% (coursework) ● Research Paper – 60% (coursework)
Assessment Details	Assessment is entirely by coursework; there is no examination. Over the course of the module students complete three programming exercises (40% in total) designed to test their ability to (1) prepare textual data for analysis, (2) apply the appropriate computational tools to extract the key quantities of interest, and (3) work with current embeddings-based models in text analysis pipelines. Each exercise is submitted separately and carries its own mark. The final assessment is a research paper (60%) of approximately 3,000 words (5–10 pages, excluding references and appendix), in which each student identifies a research question and answers it using computational text analysis tools applied to textual data.

Assessment

All work is submitted via Blackboard. Reassessment is by resubmission of the research paper (100%).

Additional Information

Prerequisites

None. Assumed knowledge: this is an intermediate-level module, and students are expected to have a basic understanding of statistics and to be comfortable with core programming concepts in R and/or Python.

Co-requisites

None

Any other information

Software: the module uses R and RStudio, both of which are free and open-source; students should install them on their own computer before the first tutorial.
Generative AI: the use of generative AI tools is permitted provided that their use is fully and transparently acknowledged, including the specific version of any model used.
Module website: <https://tom.paskhal.is/POP77032>