Quantitative Text Analysis for Social Scientists

Trinity College Dublin 2024/25

Tom Paskhalis

tom.paskhal.is

- **Module Code:** POP77142/ECP77524
- **Module Website:** tom.paskhal.is/POP77142
- **ECTS Weighting:** 5
- **Semester/Term Taught:** Semester 2 (Hillary Term)
- **Contact Hours:**
  One 2-hour lecture - Wednesday 16:00-18:00 in 4050B Arts Building
  One 2-hour tutorial:
  - Group 1 - Thursday 13:00-15:00 in PX 206 7-9 Leinster Street South
  - Group 2 - Friday 14:00-16:00 in 4053 Arts Building

  per week (5 weeks)
- **Module Coordinator:** Dr Tom Paskhalis (tom.paskhalis@tcd.ie)
- **Office Hours:** Thursday 11:00-13:00 in-person or online (booking required)
- **Teaching Fellows:**
  - Sara Cid (cids@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):

- 2 programming exercises (20% each)
- Research paper (60%)
  Approximately 5–10 pages and 3,000-4,000 words (references excluded)

The length of the final research paper provided above should serve as a guide. There is a 10% leeway in the word count. That is any submission that falls 10% short of the lower bound or exceeds by 10% the upper bound word listed above will not be penalised.

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.

2 programming exercises are due by **15:59 Wednesday** prior to the start of the lecture. See module schedule summary below for the full list of due dates.

The final research paper will be due by **23:59 Wednesday, 23 April 2025**.

Plagiarism

Plagiarism — defined by the College as the act of presenting the work of others as one's own work, without acknowledgement — is unacceptable under any circumstances. All submitted coursework must be **individual** and **original** (you should not re-use parts of a paper you

wrote for another module, for example). You need to reference any literature you use in the correct manner. This is true for use of quotations as well as summarising someone else's ideas in your own words. When in doubt, consult with the lecturer before you hand in an assignment. Plagiarism is regarded as a major offence that will have serious implications. For more information on the College policy on plagiarism, please see avoiding plagiarism guide. All students must complete the online tutorial on avoiding plagiarism which can be found on this webpage.

Module Schedule

Module Schedule Summary 6
Week 1: Introduction 6
Week 2: Quantifying Texts 7
Week 3: Classifying Texts 7
Week 4: Modelling Texts 7
Week 5: Beyond Bag-of-Words 8

Module Schedule Summary

Week	Date	Topic	Released	Due
8	12 March	Introduction	Assignment 1	
9	19 March	Quantifying Texts		
10	26 March	Classifying Texts	Assignment 2	Assignment 1
11	2 April	Modelling Texts		
12	9 April	Beyond BOW		Assignment 2

Week 1: Introduction

Required Readings:

- Justin Grimmer and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297. https://doi.org/10.1093/pan/mps028
- John D. Wilkerson and Andreu Casas. 2017. "Large-scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20:529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Additional Readings:

- Chs 1–2 Grimmer, Roberts, and Stewart 2022
- Matthew Gentzkow, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–74. https://web.stanford.edu/~gentzkow/research/text-as-data.pdf
- Kristoffer L. Nielbo et al. 2024. "Quantitative text analysis." *Nature Reviews Methods Primers* 4 (1): 25. https://doi.org/10.1038/s43586-024-00302-w

Week 2: Quantifying Texts

Required Readings:

- Ch 5 Grimmer, Roberts, and Stewart 2022
- Matthew J. Denny and Arthur Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do About It." *Political Analysis* 26 (2): 168–189. https://arthurspirling.org/documents/preprocessing.pdf

Additional Readings:

- Kasper Welbers, Wouter Van Atteveldt, and Kenneth Benoit. 2017. "Text Analysis in R." *Communication Methods and Measures* 11 (4): 245–265. https://kenbenoit.net/pdfs/text_analysis_in_R.pdf
- Paul C. Bauer, Camille Landesvatter, and Lion Behrens, eds. 2024. *APIs for social scientists: A collaborative review.* https://paulcbauer.github.io/apis_for_social_scientists_a_review

Week 3: Classifying Texts

Required Readings:

- Chs 16–20 Grimmer, Roberts, and Stewart 2022
- Pablo Barberá et al. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29 (1): 19–42. http://pablobarbera.com/static/text_practical_guide.pdf

Additional Readings:

- Lori Young and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29 (2): 205–231. https://doi.org/10.1080/10584609.2012.671234
- Kenneth Benoit et al. 2016. "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–295. https://kenbenoit.net/pdfs/Crowd_sourced_data_coding_APSR.pdf

Week 4: Modelling Texts

Required Readings:

- Chs 6, 12–13 Grimmer, Roberts, and Stewart 2022
- David M. Blei. 2012. "Probabilistic Topic Models." In *Communications of the ACM,* 55:77–84

Additional Readings:

- Kenneth Benoit, Michael Laver, and Slava Mikhaylov. 2009. "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53 (2): 495–513. https://kenbenoit.net/pdfs/blm2009ajps.pdf
- Margaret E Roberts et al. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–1082. https://scholar.harvard.edu/sites/scholar.harvard.edu/files/dtingley/files/topicmodelsopenendedexperiments.pdf

Week 5: Beyond Bag-of-Words

Required Readings:

- Chs 7–8 Grimmer, Roberts, and Stewart 2022
- Pedro L. Rodriguez and Arthur Spirling. 2022. "Word Embeddings: What works, what doesn't, and how to tell the difference for applied research." *Journal of Politics* 84 (1): 101–115. http://arthurspirling.org/documents/embed.pdf

Additional Readings:

- Ch 6 Jurafsky and Martin 2025
- Ludovic Rheault and Christopher Cochrane. 2020. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis* 28 (1): 112–133. https://lrheault.github.io/downloads/rheaultcochrane2019_pa.pdf
- Tomas Mikolov et al. 2013. "Distributed representations of words and phrases and their compositionality." In *Proceedings of the 27th International Conference on Neural Information Processing Systems,* 2:3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),* 1532–1543. https://nlp.stanford.edu/pubs/glove.pdf