Week 1: Introduction

POP77142 Quantitative Text Analysis for Social Scientists

Tom Paskhalis

Overview

- Module objectives
- Prerequisites and software
- Materials and books
- Module meetings
- Assessment and collaboration
- Weekly schedule

Module Objectives

- Introduce the fundamentals of working with text as data;
- Extract and prepare textual data for analysis;
- Apply key computational techniques for textual data;
- Practice these concepts using social science examples.

Module Materials

- Module website: tom.paskhal.is/POP77142
- Blackboard

Books

- 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.

Also:

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

Additional Online Materials

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

Prerequisites and Software

- Intermediate module - familiarity with basic statistical concepts and programming in R/Python is assumed.
- Laptop with Windows/Mac/Linux OS (no Chrome books)
- Required software:
  - **Jupyter** - web-based interactive computational environment
  - **Python** (version 3+) - versatile programming language
  - **R** (version 4+) - statistical programming language
- Additional software:
  - **JupyterLab Desktop** - desktop application for Jupyter Notebooks
  - **RStudio** - integrated development environment for R
  - **Spyder** - integrated development environment for Python
  - **Visual Studio Code** - feature-rich text editor

Module Meetings

- 2-hour lecture
  - Wednesday 16:00-18:00 in 4050B Arts Building
- 2-hour tutorials
  - 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
- Office hours:
  - Thursday 11:00 - 13:00 online or in-person (booking required)

Assessment

- 2 programming exercises (20% each)
- Research paper (60%)
  - Approximately 10 pages and 3,000-4,000 words (references excluded)
  - Due by **23:59 Wednesday, 23 April 2025**

Plagiarism

- Plagiarising computer code is as serious as plagiarising text (see Google LLC v. Oracle America, Inc.)
- All submitted programming assignments and final project should be done individually;
- You may discuss general approaches to solutions with your peers;
- But do not share or view each others code;
- You can use online resources but give credit in the comments.

Generative AI

- The use of generative AI is permitted.
- However:
  - No part of the module content can be used in a prompt;
  - It needs to be explicitly acknowledged in the submission;
  - You need to state the version of the model used.
- Hardware permitting, I recommend using local offline models installed on your machine.
- E.g. check LM Studio as a user-friendly interface to different models.

Module Outline

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

Next

- Introduction to QTA