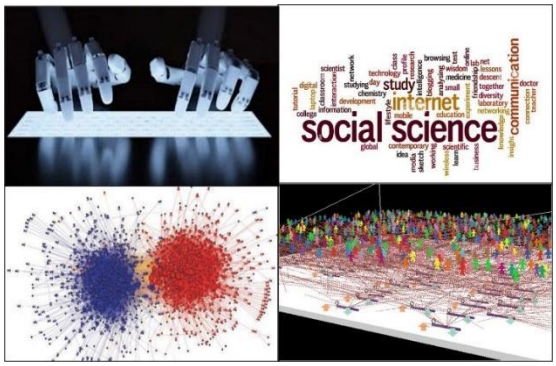
National Chengchi University, College of Social Sciences

IDAS, Course Number: 265783001

Text Analysis for the Social Sciences



Office Hours	Tu/W 2-4 PM & by appt.	**Class Time**	Tuesday 5-8pm
Office Location	綜北13樓	**Classroom #**	綜北270207
Instructor	Dr. Jacob Reidhead	**Websites**	https://moodle.nccu.edu.tw/course/view.php?id=36513
Email	reidhead@g.nccu.edu.tw		

1. Introduction

Content: The explosion of digital text represents an unprecedented opportunity for social research! This course surveys text pre-processing and analysis on three levels. Topics include:

1. **Pre-Processing**: cleaning, pre-processing, exploratory analysis
2. **Analyzing Tokens**: vectorization, word-embeddings, tagging and naming
3. **Analyzing Documents**: text classification, sentiment analysis
4. **Analyzing Corpora**: latent topics, semantic networks

Students will examine how text analysis is used to conduct social science. They will also apply text analysis to social science questions. Special attention will be given to generative AI and the ways it is rapidly augmenting text analysis.

Audience: Designed as a course for IDAS students, students in IMAS or other graduate programs are also welcome to join. No prerequisites and no prior coding experience are required. Sample

code will be provided. And as your instructor, I will walk you through each exercise, step-by-step. No fear! Let's start coding!

2. Intended Learning Outcomes

You will become familiar with:

1. Types of digital text and text analysis
2. The ways digital text and text analysis are impacting the social sciences

You will practice:

3. Implementing computational methods on sample code and text
4. Communicating your computational analyses and their broader applications
5. Using generative AI to augment your text analysis

3. Course Organization

Modules

The course is organized around four modules. Each module will introduce two or three forms of analysis, explore how AI can assist us with these analyses, and review how this analysis is used to conduct social science.

1. **Pre-Processing**: cleaning, pre-processing, exploratory analysis
2. **Analyzing Tokens**: vectorization, word-embeddings, tagging and naming
3. **Analyzing Documents**: text classification, sentiment analysis
4. **Analyzing Corpora**: latent topics, semantic networks

Teams

By the second week of the semester, the class will organize into teams of 2 or 3 students. Students will work in teams on team videos and the research poster.

Activities

- In class, students will actively participate in Lectures, Workshops and Roundtables.
- Out of class, students will work individually on Coding Tasks and Research Reviews.
- Out of class, students will work in teams on Team Videos and a Research Poster.

ILO-Activity Matrix

Course activities are designed to satisfy the following Intended Learning Outcome (ILO).

Activity	**ILO1 Learn tools and data**	**ILO2 Learn impact on social science**	**ILO3 Implement CSS methods**	**ILO4 Communicate analysis & apps**	**ILO5 Use AI to augment analysis**
1. Research Reviews		x		x	
2. Coding Tasks	x		x	x	
3. Team Videos	x	x	x	x	x
4. Research Poster	x	x	x	x	

4. Assessments & Grades

Grading Rubric

Task	**Points per Assessment**	**Percent of Semester Grade**
Attendance	After TWO FREE ABSENCES, each unexcused absence is -5%	0%
Coding	4 reviews * 5-10% each	30%
Review	4 tasks * 5- 10% each	30%
Team Videos	2 videos * 10% each	20%
Research Poster	1 poster & presentation	20%

Coding Tasks – 20%

A coding task will be assigned for Modules 1-4. The instructor will provide sample code and data. Students will execute the sample code and make small changes, as the task requires. Findings will be summarized in a 3-page report (1500 words).

Research Reviews - 20%

For each of the four modules, a research review will be assigned. The instructor will specify a computational method or type of data. Students will be asked to find three research papers using that type of method or data, and then to summarize and compare the three papers in a 3-page (1500 word) research review. Students may choose papers in a social science field which interests them.

Team Video – 40%

For Modules 2-4, students will work in teams to complete a text analysis project and make a 10-minute video presenting their results. Each project will expand on that module's coding task and research review. Students will choose a source of text reflecting a social science application of interest to them. Working in teams, students will be asked to conduct an analysis of the text. Teams may use generative AI tools in order to augment their task. Then they will create a team video, also using AI tools if they wish, to present the results of their analysis.

Research Poster & Poster Session – 20%

As an end-of-semester project, students will prepare a research poster describing their use of text analysis to answer a social science research question. Posters will have sections on background, the student's theory and hypothesis, data, analysis, discussion and conclusion. Students will submit a draft of their poster for evaluation, revise the poster based on feedback, print their final poster and present it in an end-of-semester poster session.

5. Course Materials

References

- Text Analysis
    - 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.
    - Bengfort, B., Bilbro, R., & Ojeda, T. (2018). Applied text analysis with Python: Enabling language-aware data products with machine learning. " O'Reilly Media, Inc.".
    - Szabó, G., Polatkan, G., Boykin, P. O., & Chalkiopoulos, A. (2018). Social media data mining and analytics. John Wiley & Sons.
- Machine Learning
    - Jacobucci, R., Grimm, K. J., & Zhang, Z. (2023). Machine Learning for Social and Behavioral Research. Guilford Publications.
    - Beyeler, M. (2017). Machine Learning for OpenCV. Packt Publishing Ltd.
- Social Network Analysis

- Borgatti, S. P., Everett, M. G., Johnson, J. C., & Agneessens, F. (2022). Analyzing social networks using R. Sage.
- Jackson, M. O. (2008). Social and economic networks (Vol. 3). Princeton: Princeton university press.

Moodle

All course materials are available on Moodle. Course materials include:

- Syllabus, Academic Calendar, Records of attendance and grades
- Lecture slides and Assigned Readings
- Worksheets and URLs used for all in-class activities
- Video tutorials, sample code, data and instructions for all coding tasks
- Instructions for all research reviews

6. Academic Policies

Grading Scale

The grading scale for this course follows the system typically used at NCCU.

Extra Credit

I rarely offer extra credit. However, if extra credit is offered, it will not be arbitrarily offered to individual students, but systematically offered to all students equally.

Academic Integrity

NCCU requires all students to adhere to high standards of integrity in their academic work. No type of academic misconduct (including but not limited to plagiarism, cheating, or lying to the professor) will be tolerated in this class and may result in penalties including but not limited to scores of 0 on assignments and forfeiture of extra credit points. Instances of academic misconduct will be referred directly to the appropriate disciplinary committee. For full information on these matters, please refer to the NCCU catalog or official website.

Generative AI

Students are encouraged to use generative AI to augment any aspects of all assignments including literature reviews, coding, team videos and the research poster. If AI-generated results do not fully satisfy assignment criteria, some human intervention may be required in order to complete the assignment and receive full credit.

7. Course Schedule

Week	Date	Topic	Assignment
Module 1: Pre-Processing			
1	2/20	Introduction to Text & Text Analysis	
2	2/27	Pre-Processing	
3	3/5	Keyword Analysis	
4	3/12	Analysis Workshop	Sun 3/17: Code & Review #1
Module 2: Generating Features			
5	3/19	Latent Topics	
6	3/26	Named Entities, TF-IDF, POS, Syntax Trees	Sun 3/31: Code & Review #2
7	4/2	Lexicon-Based Classification: Sentiment, etc.	Sun 4/7: Team Video
8	4/9	Module 2 Video Roundtable	
Module 3: Analyzing Documents			
9	4/16	Machine Learning Classification: Personality	
10	4/23	No Class	
11	4/30	Machine Learning Classification: Images	Sun 5/5: Code & Review #3
12	5/7	Cluster Analysis of Tokens ● Document Clustering by Tokens ● Conversation Sequence Analysis	Sun 5/12: Team Video
13	5/14	Module 3 Video Roundtable	Sun 5/19: Poster – Mock-up
Module 4: Modeling the Corpus			
14	5/21	Word Embedding Models	
15	5/28	Semantic Network Analysis ● Keyword Networks ● Named Entities Networks ● Latent Topic Networks	Sun 6/2: Code & Review #4
16	6/4	Final Roundtable ● Discuss Review #4 ● Share findings from Coding Task #4 ● Future of Text Analysis	
17	6/11	Poster Session	Tues 6/11: Poster – Printed
18	6/18	No class	Tues 6/18: Late Assignments