# Ling 583 Statistical Methods in Text Analysis
# Spring 2026
# Class # 5149

SDSU

## COURSE INFORMATION


Instructor: Prof. Rob Malouf
Email: rmalouf@sdsu.edu

Mode: Asynchronous online
Office hours: Mon/Wed 1:00–2:00 or by appointment
Real office: SHW 201
Zoom office: see Canvas

According to industry folklore, up to 80% of the information owned by large organizations is in the form of 'unstructured' text documents. During this course, students will collaboratively build applications that extract usable information from a large collection of texts. We will cover techniques for collecting, organizing and annotating textual databases. The course activities will familiarize students with a range of practical methods for analyzing large text datasets, including topic models, vector-space and embedding-based representations, logistic regression and other baseline classifiers, and transformer-based transfer learning, as applied to tasks such as topic modeling, sentiment and aspect-based analysis, text extraction, and retrieval-augmented generation (RAG) using vector databases.

Prerequisites: Ling 571 or Ling 572; and Stat 550 or Stat 551A; or permission of instructor

## LAND ACKNOWLEDGEMENT

For millennia, the Kumeyaay people have been a part of this land. This land has nourished, healed, protected and embraced them for many generations in a relationship of balance and harmony. As members of the San Diego State University community, we acknowledge this legacy. We promote this balance and harmony. We find inspiration from this land, the land of the Kumeyaay.

## ESSENTIAL STUDENT INFORMATION

For essential information about student academic success, please see the SDSU Student Academic Success Handbook.

- SDSU provides disability-related accommodations via Student Disability Services (sds@sdsu.edu|https://sds.sdsu.edu/). Please allow 10-14 business days for this process.
- Class rosters are provided to the instructor with the student's legal name. Please let me know if you would prefer an alternate name and/or gender pronoun.

## COURSE MATERIALS

All course materials will be posted on Canvas. Programming projects will be done on SDSU's instructional computing cluster (VERNE).

## COURSE DESIGN: MAJOR ASSIGNMENTS AND ASSESSMENTS

This course is organized around four projects: aspect-based sentiment analysis, text classification, retrieval augmented generation, and a final project of your design. There will also be weekly lab assignments for developing the skills you need to do the projects.

Students will work in pairs for labs (I'll assign lab partners for each unit). Projects should be done individually. After each assignment, students will write brief peer reviews of each others work.

New modules will be posted each Monday and assignments will be due the following Monday.

## COURSE SCHEDULE

| Week | Date | Subject | |
|---|---|---|---|
| 1 | 1/20 | Data exploration | |
| 2 | 1/26 | Topic models | |
| 3 | 2/2 | Parsing | |
| 4 | 2/9 | Evaluation | |
| 5 | 2/16 | Project #1 | **Aspect-Based Sentiment Analysis** |

| Week | Date | Subject | |
|---|---|---|---|
| 6 | 2/23 | Logistic regression | |
| 7 | 3/2 | Vector space models | |
| 8 | 3/9 | Transfer learning | |
| 9 | 3/16 | Error analysis | |
| 10 | 3/23 | Project #2 | **Text Classification** |
| 11 | 4/6 | Text extraction | |
| 12 | 4/13 | Vector databases | |
| 13 | 4/20 | RAG pipeline | |
| 14 | 4/27 | Project #3 | **Retrieval Augmented Generation** |

**Final project due 5/13**

## GRADING POLICIES

This course moves quickly and each week builds on what came before. It's very important not to let yourself fall behind. Since this is an asynchronous class, it's tempting to procrastinate and try to do all the labs at the end of the unit. Don't do that. Assignments are due by 8:00pm on the specified due date. By default, I will apply a **5%-per-day grade penalty for any late assignments, up to a possible 50% reduction** (it's always better to turn in something, even if very late, than to not turning anything in at all). However, I can be flexible. If problems come up (health, family, work, travel) please let me know and we can probably work out an alternate schedule.

The total grade for the class will be based on:

| | |
|---|---|
| Labs | 30% |
| Peer reviews | 5% |
| Project 1 | 15% |
| Project 2 | 15% |
| Project 3 | 15% |
| Final project | 20% |

## AI SYLLABUS STATEMENT

AI is an essential part of the text analyst's toolkit and we will be using AI extensively in this class. That said, while AI has it's place, unauthorized use of AI will be considered plagiarism and academic dishonesty. There are two very good reasons not

to use AI. One is that AI is just not very good at a lot of tasks. It can be hard to tell when it is producing superficially polished but deeply flawed output, especially for beginners. The other reason is that even for some tasks that AI is good at, offloading them to AI cheats you out of an opportunity to learn. I assume you're all here to gain skills. Some uses of AI are like lifting weights with a forklift. Sure, the weights get lifted, but ultimately it's a waste of everyone's time. For the third category – tasks which AI is good at and which aid learning rather than subverting it – AI tools can be a great help.

I will be specific in each assignment about where AI should and shouldn't be used. As a general guideline, though, you can use AI as an assistant (but not a replacement!) for:

- Locating readings and documentation
- Understanding what you are reading
- Getting your code to work
- Suggesting suitable analyses or visualizations

Do not use AI for:

- Summarizing articles that you haven't read
- Writing your code
- Writing text
- Correcting grammar or style - I want to see your ideas in your voice

If you are wondering whether a particular use of AI is legitimate, ask me.

## STUDENT LEARNING OUTCOMES

1. **Collect, preprocess, and clean text data** from diverse sources (e.g., social media, news, academic texts; APIs, databases) while considering data privacy, security, and ethical implications.
2. **Identify and apply key text pre-processing techniques** such as tokenization, term extraction, stemming, lemmatization, and stopword removal, understanding their impact on downstream analysis, and how these choices shape LLM-based vs. traditional pipelines.
3. **Implement text classification, clustering, and topic modeling** using machine learning and natural language processing (NLP) techniques in Python and rele-

vant libraries (e.g., spaCy, scikit-learn, tomotopy, huggingface), using both classical methods and transformer/LLM-based approaches, across different textual domains and genres.

4. **Apply and assess various semantic analysis techniques**, such as word embeddings, topic modeling, and semantic similarity measures, using modern embedding models and LLM-derived representations while considering real-world applications and limitations.

5. **Evaluate text mining models** based on appropriate performance metrics, ensuring transparency and reproducibility in computational analysis, with suitable evaluation of LLM outputs (e.g., reliability, bias, and consistency) when used.

6. **Design appropriate interactive visualizations and reports** to effectively communicate text mining results to both technical and non-technical audiences, clearly documenting any LLM use and its limitations.

7. **Develop reproducible text mining workflows** using version control, documentation, and best practices for computational research.

8. **Design and execute a text mining project** that integrates multiple analysis steps, applies learned techniques to a real-world dataset, and discusses ethical concerns such as data ownership, consent, and responsible

Last revised: 16 January 2026