Ling 583 Statistical Methods in Text Analysis

San Diego State University
Schedule # 8005
Spring 2024
M 4:00pm–6:40pm
SHW-243

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 the use of a range of standard statistical methods for analysis of large texts, including Markov models, Bayesian classifiers, logistic regression, support vector machines, and neural nets, as applied to tasks such as topic modeling, relation detection, and sentiment analysis. Prerequisites: Ling 571 or Ling 572; and Stat 550 or Stat 551A; or permission of instructor

Prof. Rob Malouf

Website:	malouf.sdsu.edu
Email:	rmalouf@sdsu.edu
Office hours:	TTh 1:00–2:00 or by appointment
Real office:	SHW 244
Zoom office:	SDSU.zoom.us/j/88663149131

Student Learning Outcomes

Upon successful completion, students will have the knowledge and skills to:

1. Apply techniques for collecting and preparing text data for computational analysis.
2. Describe a variety of text analysis steps, understand their interdependencies, and identify algorithms for each step.
3. Choose and apply appropriate text classification and clustering algorithms
4. Choose and apply appropriate semantic analysis techniques
5. Integrate knowledge and apply skills acquired over the sequence of text analysis courses to solve real world problems.

Course Materials

The textbooks for this course are:

- Jens Albrecht, Sidharth Ramachandran, Christian Winkler. 2020. *Blueprints for Text Analytics Using Python*. O'Reilly.
- Andreas C. Müller and Sarah Guido. 2017. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly.
- Lewis Tunstall, Leandro von Werra, & Thomas Wolf. 2022. *Natural Language Processing with Transformers (Revised edition).* O'Reilly.

All of these books are available online for free through the SDSU library. If you prefer a printed copy, they should be available on Amazon, etc. All other course information, additional readings, assignments, slides, etc. will be available on Canvas.

Course Design

The final grade will be based on in class lab assignments and three projects. Students may work in small groups, but each student will be required to submit their own write-up. As part of their final project, graduate students will also be required to write a paper describing the design of a (hypothetical) text mining application. The final grade will be 20% lab assignments, 20% each for projects #1 and 2, and 40% for the final project.

Schedule

Week	Topic
1—2	Words
3	Term extraction
4	Topic models
5	**Project #1**
6—7	Classifiers
8	Model evaluation
9	**Project #2**
10	Sentiment analysis
11—12	Deep learning
13—15	**Final projects**

Revised: 16 January 2024