Statistical Methods in Text Analysis

San Diego State University
Schedule # 22208
Spring 2022
Mo 4:00pm–6:40pm
SHW-243*

*Following campus guidelines, this class will meet synchronously on Zoom until Feb 4, 2022. After that, class will be held in person in SHW-243. Updated information will be provided via Canvas as it becomes available.

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:	Tu Th 2:00–3:00 or by appointment
Real office:	SHW 244
Zoom office:	SDSU.zoom.us/j/88612436627

Student Learning Outcomes

After completion of this course, students will be able to:

- Apply techniques for collecting and preparing text data for computational analysis.
- Describe a variety of text analysis steps, understand their interdependencies, and identify algorithms for each step.
- Choose and apply appropriate text classification and clustering algorithms
- Choose and apply appropriate semantic analysis techniques
- Use cloud computing resources to construct scalable text analysis pipelines
- 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 is:

- 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.

Both of these books are availble online for free through the SDSU library. If you prefer a printed copy, they should be available in the bookstore, on Amazon, etc. Additional readings will be made available on canvas.

Course Design

The final grade will be based short homeworks (10%), in class labs (10%), online exercises (5%) in-class labs, and three projects (25% each). Students may work in small groups, but each student will be required to submit their own write-up. Graduate students will also be required to write a paper describing the design of a (hypothetical) text mining application.

Schedule

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

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

- Compliance with CSU / SDSU vaccination and facial covering policies (newscenter.sdsu.edu/student_affairs/srr/covid-policies.aspx) is required.
- Your SDSU email address (gsuite.sdsu.edu) will be used for all course-related communications.
- The Student Conduct Code (newscenter.sdsu.edu/student_affairs/srr/conduct.aspx) prohibits conduct disruptive to instruction, including academic dishonesty and the unauthorized recording, dissemination, or publication (including on websites or social media) of lectures or other course materials.
- SDSU provides disability-related accommodations via the Student Ability Success Center (sascinfo@sdsu.edu | sdsu.edu/sasc). Please allow 10–14 business days for this process.
- The Family Educational Rights and Privacy Act (FERPA) (bfa.sdsu.edu/hr/oerc/students/ferpa.aspx) mandates the protection of student information, including contact information, grades, and graded assignments. I will not post grades or leave graded assignments in public places. Students will be notified at the time of an assignment if copies of student work will be retained beyond the end of the semester or used as examples for future students or the wider public.
- As an instructor, one of my responsibilities is to help create a safe learning environment on our campus. I am required to share information regarding sexual violence on SDSU's campus with the Title IX (titleix.sdsu.edu) coordinator, Gail Mendez (619-594-6464), who will contact you to let you know about support services at SDSU and possibilities for holding accountable the person who harmed you. If you do not want the Title IX Officer notified, you can speak confidentially SDSU's Sexual Violence Victim Advocate (619-594-0210) or Counseling and Psychological Services (619-594-5220, psycserv@sdsu.edu).
- 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.
- Need help finding an advisor, tutor, counselor, emergency economic assistance, or other support? Contact the SDSU Student Success Help Desk (studentsuccess.sdsu.edu) Monday through Friday, 9:00 AM to 4:30 PM.
- For technical or computing assistance, contact the Library Computing Hub (virtual-academic-help.sdsu.edu/technology).

Absences

- According to the University Policy File, students should notify the instructors of affected courses of planned absences for religious observances by the end of the second week of classes.
- If you are absent more than five days due to illness or injury, you may contact Student Health Services (shs.sdsu.edu/index.asp) for help in communicating your absence.
- If you miss class because you have been diagnosed with or are required to quarantine due to exposure to COVID-19, contact vpsafrontdesk@sdsu.edu to notify the university

Academic Honesty

The University adheres to a strict policy prohibiting cheating and plagiarism, including

- Copying, in part or in whole, from another's test or other examination.
- Obtaining copies of a test, an examination, or other course material without the permission of the instructor.
- Collaborating with another or others in coursework without the permission of the instructor.
- Falsifying records, laboratory work, or other course data.
- Submitting work previously presented in another course, if contrary to the policies of the course.
- Altering or interfering with grading procedures.
- Assisting another student in any of the above.
- Using sources verbatim or paraphrasing without giving proper attribution (this can include phrases, sentences, paragraphs and/or pages of work).
- Copying and pasting work from an online or offline source directly and calling it one's own.
- Using information found from an online or offline source without giving the author credit.
- Replacing words or phrases from another source and inserting one's own words or phrases.

Under CSU policy, instructors must report instances of academic misconduct to the Center for Student Rights and Responsibilities for disciplinary review by the University, which may lead to probation, suspension, or expulsion. Instructors may also, at their discretion, penalize student grades on any assignment or assessment discovered to have been produced in an academically dishonest manner.

Revised: 12 January 2022