Ling 583 Statistical Methods in Text Analysis

Schedule #22131
Spring 2021
M 4:00pm–6:40pm
On-line: SDSU.zoom.us/j/87963680553

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: Tue & Wed 1:00pm–2:00pm, or by appointment
Zoom office: SDSU.zoom.us/j/81821137566
Google Chat room: chat.google.com/room/AAAAxC2xL_M

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.

Communications

We'll be using several different modes of communication this semester. My goal is to get as little email as possible, so I'd prefer to use other communication channels as appropriate. Official class communications, like announcements, assignments, videos, etc., will be posted on **Canvas**. To access your course log in at `canvas.sdsu.edu` and sign in using your SDSUid. You can also contact me or other students using the Canvas 'Conversations' function: click on the Inbox icon and go from there.

Note: You are responsible for adjusting your notification settings in such a way that you receive all announcements regarding this class. All Canvas email notifications will be delivered to your SDSU email address. You can add additional email addresses and sign up for text/mobile app notifications via the settings in your Canvas Profile, and then adjust your notifications in the Notifications Tab. If you have technical issues with Canvas, please contact the SDSU Canvas 24/7 support line at 619.483.0632.

Synchronous communications, like lectures and office hours, will be run using **Zoom** for video and in-class chat. All Zoom class sessions (but not office hours) will be recorded and posted to Canvas afterwards, so even if you aren't able to attend you'll still be able to see what you missed. For details about Zoom meetings, see Canvas.

For informal conversations, we'll use **Google Chat**. This is where you can ask questions about readings, set up study groups, chat with other students, and so on. I'll be available on Google Chat, but nothing that happens there will be graded. I will send out invitations for the Google Chat room to all registered students the first week of class (let me know if you didn't get one).

Finally, if you need to you can always send me **email** at `rmalouf@sdsu.edu`. Having all classes fully online is going to be a new experience for a lot of us, and my hope is that having multiple channels will make things easier and help support a sense of community. It's not supposed to make things harder, and there's **no wrong way** to get in touch with me. If you can't figure out which mode is best, just send me an email: that will always work, and I won't hold it against you!

No matter what channels you use, please be respectful in your interactions with your fellow students (see the 'Classroom conduct' and 'Prohibited harassment' sections below) and let someone know about any problems that come up (see 'Student affairs').

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

Due to the ongoing pandemic, this class will be entirely online. Lectures, assignments, exams, and readings will either be posted to Canvas or held via Zoom. Any Zoom lectures will be recorded and posted to Canvas: synchronous attendance is optional (though highly recommended).

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.

University Policies

Accommodations

If you are a student with a disability and are in need of accommodations for this class, please contact Student Ability Success Center at 619-594-6473 as soon as possible. Please know accommodations are not retroactive, and I cannot provide accommodations based upon disability until I have received an accommodation letter from Student Ability Success Center.

Student privacy and intellectual property

The Family Educational Rights and Privacy Act (FERPA) mandates the protection of student information, including contact information, grades, and graded assignments. I will use Canvas to communicate with you, and 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. Students maintain intellectual property rights to work products they create as part of this course unless they are formally notified otherwise.

Religious observances

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.

Deferred Action for Childhood Arrivals (DACA)

All students with questions or concerns regarding DACA are encouraged to see the College of Health and Human Services Assistant Dean for Student Affairs, Jessica Robinson (`jmrobinson @sdsu.edu`).

Academic honesty

The University adheres to a strict policy prohibiting cheating and plagiarism. Examples of academic dishonesty include but are not limited to:

- 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 work to be presented 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 rules 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 your own;

- using information you find from an online or offline source without giving the author credit;

- replacing words or phrases from another source and inserting your own words or phrases.

The California State University system requires instructors to report all instances of academic misconduct to the Center for Student Rights and Responsibilities. Academic dishonesty will result in disciplinary review by the University and 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.

Resources for students

A complete list of all academic support services–including the Writing Center and Math Learning Center–is available on the Student Affairs' Academic Success website. Counseling and Psychological Services (619-594-5220) offers confidential counseling services by licensed therapists; you can Live Chat with a counselor at `go.sdsu.edu/student_affairs/cps/therapist-consultation.aspx` between 4:00pm and 10:00pm, or call San Diego Access and Crisis 24-hour Hotline at (888) 724-7240.

Sexual violence / TItle IX mandated reporting

As an instructor, one of my responsibilities is to help create a safe learning environment on our campus. I am a mandated reporter in my role as an SDSU employee. It is my goal that you feel able to share information related to your life experiences in classroom discussions, in your written work, and in our one-on-one meetings. I will seek to keep the information you share private to the greatest extent possible. However, I am required to share information regarding sexual violence on SDSU's campus with the Title IX coordinator, Jessica Rentto 619-594-6017. She (or her designee) will contact you to let you know about accommodations and support services at SDSU and possibilities for holding accountable the person who harmed you. Know that you will not be forced to share information you do not wish to disclose and your level of involvement will be your choice. If you do not want the Title IX Officer notified, instead of disclosing this information to your instructor, you can speak confidentially with the following people on campus and in the community. They can connect you with support services and discuss options for pursuing a University or criminal investigation. Sexual Violence Victim Advocate 619-594-0210 or Counseling and Psychological Services 619-594-5220, `psycserv@sdsu.edu`. For more information regarding your university rights and options as a survivor of sexual misconduct or sexual violence, please visit `titleix.sdsu.edu` or `sdsutalks.sdsu.edu`.

Classroom conduct standards

SDSU students are expected to abide by the terms of the Student Conduct Code in classrooms and other instructional settings. Prohibited conduct includes:

- Willful, material and substantial disruption or obstruction of a University-related activity, or any on-campus activity.

- Participating in an activity that substantially and materially disrupts the normal operations of the University or infringes on the rights of members of the University community.

- Unauthorized recording, dissemination, or publication (including on websites or social media) of lectures or other course materials. This includes posting of exam problems or questions to on-line platforms.

- Conduct that threatens or endangers the health or safety of any person within or related to the University community, including physical abuse, threats, intimidation, or harassment; or sexual misconduct.

Violation of these standards will result in referral to appropriate campus authorities.

Medical-related absences

Students are instructed to contact their professor/instructor/coach in the event they need to miss class, etc. due to an illness, injury or emergency. All decisions about the impact of an absence, as well as any arrangements for making up work, rest with the instructors. Student Health Services (SHS) does not provide medical excuses for short-term absences due to illness or injury. When a medical-related absence persists beyond five days, SHS will work with students to provide appropriate documentation. When a student is hospitalized or has a serious, ongoing illness or injury, SHS will, at the student's request and with the student's consent, communicate with the student's instructors via the Vice President for Student Affairs and may communicate with the student's Assistant Dean and/or the Student Ability Success Center.

SDSU Economic Crisis Response Team

If you or a friend are experiencing food or housing insecurity, or any unforeseen financial crisis, visit sdsu.edu/ecrt, email ecrt@sdsu.edu, or walk-in to Well-being & Health Promotion on the 3rd floor of Calpulli Center.

Concerns regarding classroom activity, grades, or other student affairs matters

Though students have the option of contacting the San Diego State University Ombudsman or the Assistant Dean for Student Affairs at any time regarding classroom activity, grades, or other student affairs matters, students are encouraged to meet with their professor first to discuss the situation. If the issue is not resolved at this level, the student should contact their undergraduate or graduate advisor. If the problem is not resolved at this level, contact should be made to their department director and finally their Assistant Dean for Student Affairs.

Prohibited harassment

The CSU has affirmed its commitment to "protecting access, affordability, intellectual freedom, inclusivity, and diversity for all students ... including supporting DACA students." Discrimination, harassment, or retaliation against students, faculty, and staff on the basis of race, religion, gender, sexuality, disability, nationality, immigration status and other categories of identity is prohibited. If you have concerns about your status at the university, please visit `studentaffairs.sdsu.edu/EOP` for information or contact the Dean of Students or the Assistant Dean for Student Affairs in your College.

Revised: 18 December 2021