ISE 540 Text Analytics
Units: 4.0
Two 110-minutes lectures per week
M/W 12-1:50 pm, **Location:** KDC 235

Instructor: Mayank Kejriwal
Office: USC Information Sciences Institute
Office Hours: By appointment
Contact Info: kejriwal@isi.edu

Teaching Assistant: Yidan Sun
Office Hours / Office (beginning Week 2): Tuesday 2-3 pm (OHE 310S); Thursday 10:30-11:30 am (zoom, link to be shared on blackboard)
Contact Info: yidans@isi.edu

Catalog Description

Methods and algorithms for automated text analysis; machine learning; predictive web data analytics; information retrieval; social media data; natural language documents and graphs.

Course Overview

This course focuses on foundations, techniques, applications and algorithms for conducting predictive analytics on problems that involve significant text data, including webpages, social media, 'natural language' documents and even graphs. Students will learn the practical aspects of the techniques needed to build predictive analytical systems over text data. Today, many of these systems are applications of machine learning, including supervised and unsupervised learning. Topics include information retrieval (including search and indexing), natural language processing (including information extraction and entity linking), and knowledge discovery. The class will be run as a fast-paced lecture course with lots of student participation and significant hands-on experience. As an integral part of the course each student will do a project using the research and tools covered in the class. The class will occasionally feature guest lecturers with advanced knowledge in some of the covered topical areas.

Learning Objectives and Outcomes

After completing this course, students should be able to:

- Identify the fundamentals and limitations of building predictive analytics systems for real-world problems involving text data
- Explain the different aspects of text data (including structured and unstructured data, proprietary and public data, and social media data) from the lens of Big Data (4 Vs of volume, veracity, velocity and variety);
- Identify the different components in a predictive analytics ecosystem, including differences in input data (e.g., website vs. social media), evaluation metrics, cloud and infrastructure, and algorithmic tradeoffs;
- Describe both theory and practice in doing predictive analytics on text data,
- Apply course techniques to an actual project designed in a team setting;

- Structure a text analytics problem, and reason about the validity, utility and tradeoffs of competing solutions in real-world settings

Prerequisite(s): None

Recommended Preparation: Knowledge of a programming language on the level of ISE 150; undergraduate statistics on the level of ISE 225; ISE 529 Predictive Analytics is highly recommended.

Course Notes

The course will be run as a lecture class with student participation strongly encouraged. The first 2-3 weeks of the course are structured as a quickstart to provide a primer on fundamentals, followed by deeper presentations and more technical material for the remainder of the course. Note that this is not an engineering data analytics course: we will not be going into depth into the theory and math of machine learning or statistics. Students will be expected to review relevant aspects of such material (I will post regular and accessible pointers) before coming to class. There will be weekly readings and students are encouraged to do the readings prior to the discussion in class. All of the course materials, including lecture slides and homeworks will be posted online on blackboard. The class project is a significant aspect of this course and at the end of the semester students will present their projects in class.

Technological Proficiency and Hardware/Software Required

All assignments and lectures will assume electronic access to blackboard. Programming assignments will be in Python using Jupyter notebooks, all of which are freely available.

Required Readings and Supplementary Materials

There is no required textbook. I will be posting all relevant material online on blackboard.

Description and Assessment of Assignments

Homework Assignments

There will be **six homework assignments** for the first 11 weeks of class. The assignments must be done individually. The homework assignments are expected to take 8-10 hours per week; some will involve programming. Each assignment is graded on a scale of 0-100 and the specific rubric for each assignment is given in the assignment. You must use Jupyter notebooks (we will usually provide templates) for your assignments.

Course Project

An integral part of this course is the course project, which builds on the topics and techniques covered in the class. Students can work in teams of 4-5 people on this project (we may allow six under rare circumstances). They will write their project proposals, conduct experiments, and develop their poster as a team. We will likely have a 'poster' session in the final weeks of class for you to present your project as a group in an interactive fashion. It is my intention to have guest 'judges' on the day of the poster presentation to provide feedback and comments.

Project description: Each project team will build a text analytics application for a topic of their choice. The application will be based on real-world text datasets. We will give you good options for text datasets, from which you must select your dataset and do the project. Your team's project can (and almost certainly will) rely on publicly available codebases and platforms, but the final system should be an original analytics application. During early phases of the project, I will expect you to identify the dataset you will use, the problem you intend to solve, and measurable outcomes you expect from your solution. I will point you to other relevant data and software resources if necessary. The best projects tend to build on many of the topics covered in the class. Some questions to think about (although it is unlikely you will be able to address all of these in a single project) when devising your problem statement include: Why does anyone care about your problem, and why is it a predictive analytics problem? What are you measuring, and how? How would you validate your methods (i.e. what are your metrics and key performance indicators)? What are the biases in data that might impede your progress? How can you prove your method would generalize beyond a single case? How can you best visualize your results?

The grading breakdown of the project will be released ahead of time on blackboard. Generally, the proposal will constitute 15% (of the project grade), the poster presentation will be 70%. You will also be peer-reviewing each other's posters, details of which will be released as we get closer to the session. The peer review will constitute the last 15% of the project grade. Overall, the project will contribute to 30% of your *final* grade (see below).

Quizzes: Quizzes will always be based on the material covered in the last two class days + readings. The lowest quiz grade will be dropped. Missed quizzes will receive a zero grade, and there will be no make-up quizzes for any reason. Quizzes will not be held every week. The first quiz will be in week 2. In total, quizzes count for 8% of your grade.

Midterm: There is no mid-term for this class.

Attendance: We will take attendance randomly in some classes, where we will pass around a sheet and ask you to sign your name. If you intend to be absent for a class, or be present only on zoom, you must give us advance notice with a valid reason, otherwise your absence will not be excused excepting an emergency. We will start taking attendance starting from the second week, as we realize the first week is chaotic. We will drop two absences during the semester, but otherwise, the attendance will count for 2% of your total grade.

Final Exam: The comprehensive final exam is open book. The final exam is on the date designated by USC. If a student is unable to take it in the designated date, he/she must reach out to the instructor, well in advance. Otherwise the student must have a documented emergency at the time. Do not book flight dates that are before the final; we will not be making exceptions unless there is a documented emergency!

Grading Breakdown

Assignment	**Points**	**% of Grade**
Quizzes	11 total quizzes*10 points each (lowest quiz will be dropped) = 100	8
Attendance	Randomly taken during class (we will drop up to two absences)	2
Homework	50 each*6=300	30
Final	300	30
Class project	300	30
TOTAL	1000	100

Continued on next page

Course guideline and policies

Assignment Submission Policy

Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. You can submit homework up to one week late, but you will lose 25% of the possible points for the assignment. After one week, the assignment cannot be submitted.

Additional Policies

It is my expectation that students make every effort to attend every class, and quizzes will be designed to enforce this policy. There will also be a strict no-cellphone policy. Readings for each class are posted below as links. Students are encouraged to do these readings **before** coming to class. These readings may prove useful to you as you navigate your career in today's competitive economy, and are generally from industrial sources that will help you be informed on subject matter. Occasionally, quizzes will be given at the beginning of class and may involve the readings for that class day as test material.

Communication and Blackboard:

Blackboard will be my primary method of communicating with you. Along with course materials, I will post any syllabus updates and information about class sessions, including preparation requirements. E-mails sent to the class originate from the Blackboard system. It is your responsibility to check Blackboard daily for any new information posted relevant to upcoming sessions. We will also attempt to send announcements when we post something new, but you should still make it a point to check blackboard at least once daily.

Please be sure your e-mail address and account settings in Blackboard are correct and that you are able to receive messages from Blackboard etc.

Technology Policy:

Please do not use personal communication devices, such as cell phones, during class. Students' videotaping of faculty lectures is not permitted due to copyright infringement regulations. Use of any recorded or distributed material is reserved exclusively for the USC students registered in this class.

ChatGPT/Generative Models Policy:

While you are allowed to use ChatGPT (except in examinations), you **must** attribute it and provide the prompt(s) that you entered as an appendix to the homework. If you don't, and your answer matches with someone else's, you may be cited for plagiarism and receive a 0 on the entire homework. Furthermore, **you** will be responsible for all outputs that ChatGPT produces and that you decide to submit in your homework. It is not an excuse to assume that the answer provided by ChatGPT is always correct or complete.

No Recording and Copyright Notice:

It is a violation of USC's Academic Integrity Policies to share course materials with others without permission. No student may record any lecture, class discussion or meeting without prior express written permission. The word "record" or the act of recording includes, but is not limited to, any and all means by which sound or visual images can be stored, duplicated or retransmitted whether by an electro- mechanical, analog, digital, wire, electronic or other device or any other means of signal encoding. I reserve all rights, including copyright, to my lectures, course syllabi and related materials, including summaries, PowerPoints, prior exams, answer keys, and all supplementary course materials available to the students enrolled in my class whether posted on BB or otherwise. They may not be reproduced, distributed, copied, or disseminated in any media or in any form, including but not limited to all course note-sharing websites. Exceptions are made for students who have made prior arrangements with DSP and me.

Retention of Graded Coursework:
Final projects and any other graded work which affected the course grade will be retained for one year after the end of the course if the graded work has not been returned to the student.

Continued on next page

Course Schedule: A Weekly Breakdown

	Topics/Daily Activities	Optional Readings
Week 1	**Introduction to Course and Overview of Syllabus** **Background and Motivation:** What is predictive analytics? What are some examples? Why is text so important? **Statistics:** Overview, review of key concepts **Probability and Statistical Significance:** Applications to practical analytics problems	**None**
Week 2	**Types of Text Data:** Web, social media, natural language **Primer on Artificial Intelligence and Machine Learning:** What is 'AI' and what are the key components? Is AI the same as machine and deep learning? Understanding the connection between learning and generalization **Supervised Classification:** introduction to workflow, and examples of supervised classification problems	Text analytics on Microsoft Azure
Week 3	**Text Classification:** Real-world applications, standard workflow, feature engineering **Text Classification Cont'd:** tf-idf, simple vector space models **Representation Learning:** word embedding models, including word2vec, fastText and BERT	**Reading:** A Tour of Machine Learning Algorithms Text Classification and Naïve Bayes Text Classification Algorithms: A Survey (Sections 1 and 7 are compulsory, but I encourage you to skim through the rest)
Week 4	**Unsupervised Machine Learning:** Clustering **Clustering (Cont'd):** Different link functions, brief review of differences between clustering algorithms and measuring quality of clustering	**Reading:** String similarity (Sections 2.1 and 2.2 of dissertation, and all subsections within)
Week 5	**Pairwise Problems:** String matching and name matching and deep look at the edit distance algorithm, and its complexity and limitations **Information Retrieval (IR):** The anatomy of a search engine, indexing and evaluation of IR	**Reading:** Scoring, term weighting and the vector space model (Sections 6.2 and 6.3, and all subsections within)

Week 6	**Information Retrieval Cont'd** **Natural Language Processing (NLP)**: What is it and why is it hard? Understanding the differences between syntax, semantics and pragmatics	**Reading:** 5 Amazing Examples of NLP in Practice
Week 7	**Problems in NLP:** Information Extraction (IE), Word Sense Disambiguation (WSD), Semantic Role Labeling	**Project proposals due**
Week 8	**Intro to Advanced Problems in NLP:** Summarization, Question Answering **Big Data:** 4Vs and relevance to analytics today, introduction to MapReduce **Deep look at K-Means in MapReduce**	**Reading:** Big Data: What it is and why it matters
Week 9	**From text to graphs:** understanding structure in complex data through network science **Knowledge Graphs:** From text and/or networks to Knowledge Graphs Example Application: Google Knowledge Graph	**Reading:** Things, not strings
Week 10	**Web and AI:** Semantic Web and Linked Data I **Knowledge Graph Identification:** techniques for Entity Resolution and its evaluation, knowledge graph embeddings (briefly) **Applications:** link prediction, recommendations	**Video:** Tim Berners-Lee on the Semantic Web
Week 11	**Semantic Web and Linked Data II:** Semantic Web Layer Cake and Linked Data principles	
Week 12	Advanced topics/guest lecture; **Web and AI cont'd** **Ontologies and Web Ontology Language**	**Reading:** Industry-scale Knowledge Graphs: Lessons and Challenges (make sure to download the read the full article)
Week 13	**Applications of Knowledge Graphs:** Government, Science and Non-Profit **Examples:** KGs for COVID-19, KGs for e-commerce	
Week 14	**Closing thoughts and Applications of KGs in enterprise**	
Week 15	**Project poster presentations**	
Week 16	**Project peer-reviews due**	

FINAL	Refer to the final exam schedule in the USC *Schedule of Classes* at classes.usc.edu.	

Statement on Academic Conduct and Support Systems

Academic Integrity:
The University of Southern California is a learning community committed to developing successful scholars and researchers dedicated to the pursuit of knowledge and the dissemination of ideas. Academic misconduct, which includes any act of dishonesty in the production or submission of academic work, comprises the integrity of the person who commits the act and can impugn the perceived integrity of the entire university community. It stands in opposition to the university's mission to research, educate, and contribute productively to our community and the world.

All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s).

Other violations of academic integrity include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), collusion, knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university. All incidences of academic misconduct will be reported to the Office of Academic Integrity and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.

Students and Disability Accommodations:

USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline is comprised of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services (though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost’s Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu

Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.