ISE 540
Text Analytics
Units: 3.0
Fall 2020
Monday & Wednesday 2-3:20PM

Location: Virtual

Instructor: Professor Mayank Kejriwal
Office: USC Information Sciences Institute
Office Hours: After each class, or by appointment
Contact Info: **kejriwal@isi.edu**

Catalogue Course Description
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. Topics include applied natural language processing, information retrieval and semantic web.

Expanded Course Description
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
The learning objectives for this course are:

- Understand the fundamentals and limitations of building predictive analytics systems for real-world problems involving text data;
- Understand 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);
- Understand 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;
- Gain an appreciation of both theory and practice in doing predictive analytics on text data, and apply course techniques to an actual project designed in a team setting;
- Understand how to structure a text analytics problem, and reason about the validity, utility and tradeoffs of competing solutions in real-world settings

Prerequisite(s): An undergraduate-level course on statistics is a minimum prerequisite, since we will be regularly relying on statistical methods like significance testing, normal distributions etc.

Recommended Preparation: Knowledge of a programming language such as R or Python is desirable, some background in predictive analytics and AI . An Engineering Data Analytics course like ISE 529 is highly recommended but not required. Unless an exception is sought with good reason, we will use Python as the programming language for assignments.

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, which is 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 **bi-weekly 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.

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 2-3 people on this project. They will present their project proposals in class, conduct the project, and then present the project in class. A short, written project report will also be due upon project completion. It is my intention to have guest 'judges' on the day of the project presentation to provide feedback and comments.

Project Timeline:

- Week 8: Project proposals presented in class (team members, topic)
- Week 11: Project status update due (at most 1 page status report; format will be released on blackboard)
- Week 15: Project presentation in class (short talk) + written report (format to be released on blackboard)

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 data that is either publicly available, or can be collected and used for academic work from a public resource (e.g., the Twitter API). The application 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 datasets you are using, your collection methodology (if you're collecting the data) and the software resources that will help you achieve your goal. I will point you to relevant data and software resources if necessary. The best projects tend to build on many of the topics covered in the class. Questions to think about 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 collection? How can you prove your method would generalize beyond a single crisis? 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 10% (of the project grade), the update(s) will be 5%, the written report will be 35%, and the presentation will be 50%. Overall, the project will contribute to 30% of your final grade (see below)

Grading Breakdown

Quizzes: Quizzes will always be based on the material covered in the last two class dayas + 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. I will make the quiz available online *during* class. Quizzes may be open or closed book. Quizzes will not be held in every class, and I reserve the right to give no advance notice for a quiz. Hence, students should strive to attend every class, and seek permission in advance if they plan to miss a class. If you must attend lecture asynchronously, please reach out to me in advance with your reasons and I will make every effort to accommodate you.
Midterm: There is no mid-term for this class.
Homework: There will be bi-weekly homeworks.
Final Exam: There is a final exam at the end of the semester covering all of the material covered in the class. The final exam will be on the date designated by USC
Class Project: Each student will do a group class project based on the topics covered in the class. Students will propose their own project, do the research, write a report and present the project in class.

Assignment	**Points**	**% of Grade**
Quizzes	11 total quizzes*10 points each (lowest quiz will be dropped) = 100	10
Homework	50 each*6=300	30
Final	300	30
Class project	300	30
TOTAL	1000	100

Grading Scale

Course final grades will be determined using the following scale

A	95-100
A-	90-94
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
D+	67-69
D	63-66
D-	60-62
F	59 and below

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.

Grading Timeline

Homeworks will be returned, with feedback, the week after submission. Homework and quiz solutions will be released soon after the homework submission, or quiz, date.

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. Since the class is virtual this fall, additional course guidelines are noted on the next page. Readings for each class are posted below as links. Students must do these readings **before** coming to class. These readings are particularly important 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.

Course Schedule: A Weekly Breakdown

	Topics/Daily Activities	**Deliverables/Releases/ Readings**
Week 1 Aug. 24	**Introduction to Course and Overview of Syllabus** **Background and Motivation:** What is predictive analytics? What are some examples? Why is text so important? **Probability and Statistics:** Overview, review of key concepts **Big Data:** 4Vs and relevance to analytics today	**None**
Week 2 Aug. 31	**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? Supervised Classification	**HW1 released** **Reading:** Big Data: What it is and why it matters Text analytics on Microsoft Azure
Sept. 7	**Labor Day: No Class**	
Week 3 Sep. 9	**Machine Learning Cont'd:** Unsupervised Methods (clustering) **Text Classification:** Real-world applications, standard workflow, feature engineering	**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 Sep. 14	**Text Classification Cont'd:** tf-idf, simple vector space models, word embeddings (basics only) **Pairwise Problems:** String matching, name matching and and entity resolution	**HW1 due /HW2 released** **Reading:** String similarity (Sections 2.1 and 2.2 of dissertation, and all subsections within)
Week 5 Sep. 21	**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 Sep. 28	**Information Retrieval Cont'd** **Natural Language Processing (NLP)**: What is it and why is it hard?	**HW2 due/ HW3 released** **Reading:** 5 Amazing Examples of NLP in Practice
Week 7 Oct. 5	**Problems in NLP:** Information Extraction (IE), Word Sense Disambiguation (WSD) **Oct. 7:** Guest lecture on knowledge graphs	**Project proposals due**

Week 8 **Oct. 12**	**Problems in NLP Cont'd** **Knowledge Graphs:** From text and/or networks to Knowledge Graphs Example Application: Google Knowledge Graph	**HW3 due** **Reading:** Things, not strings
Week 9 **Oct. 19**	Knowledge Graph Identification: Entity Resolution, knowledge graph embeddings (briefly) **Applications:** link prediction	**HW4 released** **Video:** Tim Berners-Lee on the Semantic Web
Week 10 **Oct. 26**	**Web and AI:** Semantic Web, Knowledge Graphs and Linked Data **Applications of Knowledge Graphs** in Industry, Science and Non-Profit Example: KGs for COVID-19	**Reading:** Industry-scale Knowledge Graphs: Lessons and Challenges (make sure to download the read the full article)
Week 12 **Nov. 2**	**Project status update presentations**	**HW 4 due/HW5 released**
Week 13 **Nov. 9**	Advanced topics/guest lecture	
Week 14 **Nov. 16**	**Student presentations** **Course wrap-up:** Where is text analytics headed?	**HW5 due/ project presentations due**
Week 15 **Nov. 23**	**Course recap and final overview**	**Classes end Nov. 24/ Project written report due**
FINAL	Friday, Dec. 4: 2-4pm. I will make the final open book. I highly recommend taking the final during this time slot but if you are unable, you must reach out to me well in advance unless you have a documented emergency at the time.	

Additional Course Guidelines

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.

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.

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.

Statement on Academic Conduct and Support Systems

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct.

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the Office of Equity and Diversity http://equity.usc.edu or to the Department of Public Safety http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us. This is important for the safety of the whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. The Center for Women and Men http://www.usc.edu/student-affairs/cwm/ provides 24/7 confidential support, and the sexual assault resource center webpage http://sarc.usc.edu describes reporting options and other resources.

Support Systems:

Student Health Counseling Services - (213) 740-7711 – 24/7 on call
engemannshc.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-4900 – 24/7 on call
engemannshc.usc.edu/rsvp
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED) | Title IX - (213) 740-5086
equity.usc.edu, titleix.usc.edu
Information about how to get help or help a survivor of harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants. The

university prohibits discrimination or harassment based on the following protected characteristics: race, color, national origin, ancestry, religion, sex, gender, gender identity, gender expression, sexual orientation, age, physical disability, medical condition, mental disability, marital status, pregnancy, veteran status, genetic information, and any other characteristic which may be specified in applicable laws and governmental regulations.

Bias Assessment Response and Support - (213) 740-2421
studentaffairs.usc.edu/bias-assessment-response-support
Avenue to report incidents of bias, hate crimes, and microaggressions for appropriate investigation and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Support and Advocacy - (213) 821-4710
studentaffairs.usc.edu/ssa
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.
Diversity at USC - (213) 740-2101
diversity.usc.edu
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
dps.usc.edu, emergency.usc.edu
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) 442-120 – 24/7 on call
dps.usc.edu
Non-emergency assistance or information.