UCI Paul Merage School of Business

BANA 290

Advanced Data Analytics for Natural Language Processing

Spring 2018

Conal Sathi

Sameer Singh

https://canvas.eee.uci.edu/courses/9097

Course Overview

This course will cover applications of machine learning methods, such as supervised learning classifiers and unsupervised algorithms, to create computer-based models to understand and analyze text data, e.g. for text categorization, sentiment analysis, clustering documents based on similarity, etc. We will explore these methods on numerous data sets from industry.

Prerequisite: Knowledge of Python programming is required for this course.

Contact Information

Conal Sathi, Lecturer
Paul Merage School of Business
Email: csathi@uci.edu
Office Hours: By Appointment

Dr. Sameer Singh, Assistant Professor
Department of Computer Science
4204 Bren Hall
Email: sameer@uci.edu
Office Hours: By Appointment (see https://calendly.com/sameersingh/office-hours)

Professor Bio

Conal Sathi completed his master's degree at Stanford University studying computer science with the focus in Artificial Intelligence. At Stanford, he taught a graduate level Natural Language Processing (NLP) course and published research in the intersection of NLP and Graph Mining. After Stanford, he was hired as the first machine learning engineer at a technology startup focused on e-commerce analytics, which was later acquired by a large global corporation Rakuten. During his time there, he built machine learning engines for dealing with unstructured text data (from research to a production setting dealing with billions of documents). He grew and managed a team of data scientists and engineers, presented at a number of data conferences, and wrote several patents for the text classification engines he designed at the company. Currently, he is an advisor and investor to several technology companies in Northern California.

Dr. Sameer Singh is an Assistant Professor of Computer Science at the University of California, Irvine. His research focuses on large-scale and interpretable machine learning applied to information extraction and natural language processing. Before UCI, Sameer was a Postdoctoral Research Associate at the University of Washington. He received his PhD from the University of Massachusetts, Amherst in 2014, during which he also interned at Microsoft Research, Google Research, and Yahoo! Labs. He was awarded the Adobe Research Data Science Faculty Award, was selected as a DARPA Riser, won the grand prize in the Yelp dataset challenge, and received the Yahoo! Key Scientific Challenges fellowship. Sameer has published extensively at top-tier machine learning and natural language processing conferences.

Classroom Etiquette, Guidelines, & Policies

Academic Honesty
By enrolling in this course, you agree to be bound by the University of California, Irvine's policy on academic honesty (http://www.senate.uci.edu/senateweb/default2.asp?active_page_id=754). This policy may also be found in your Graduate Student Handbook.

Attendance
Your attendance for each class session is expected, as is your active participation. If you miss a class for personal or business reasons, please inform the instructor in advance if at all possible. Absences without pressing reasons indicate disinterest in the course and will reflect on your grade. Initiate arrangements for submitting any make-up assignments.

Diversity & Inclusiveness Policy
The University of California, in accordance with applicable Federal and State law and University policy, does not discriminate on the basis of race, color, national origin, religion, sex, gender identity, pregnancy, physical or mental disability, medical condition (cancer related or genetic characteristics), ancestry, marital status, age, sexual orientation, citizenship, or service in the uniformed services. The University also prohibits sexual harassment. This nondiscrimination policy covers admission, access, and treatment in University programs and activities.

Course Materials

Recommended:

- Jurafsky and Martin: Speech and Language Processing (2nd edition), Prentice Hall (2008), Chapters from the 3rd edition available for free online.

Grading

Homework Assignments	40%
Final Project	40%
Final Project Peer Reviews	5%
Class Participation	15%
TOTAL	**100%**

Homework Assignments (40%)
There will be 4 homework assignments throughout the quarter. These homework assignments will be a mix of questions to be answered and practical programming assignments dealing with real-world data. Homework assignments must be done individually.

Final Project (40%)
This will be a team project doing significant analytics and machine learning on a real-world data set. The grading for the project will be broken up into three sections:

Proposal: 5%
Presentation: 15%
Report: 20%

Final Project Peer Reviews (5%)
On the final class, each project team will present their final project. Each student will submit a write up of their peers' projects.

Class Participation (15%)
Students are expected to attend all classes, participate regularly in the large class discussion, and small group discussions. Part of this course involves working with others and seeking feedback from your peers on your in-class exercises. A large portion of the class requires you to actually work with data and apply some of the learnings from the lectures and readings on that data.

IMPORTANT: You must have access to a computer on which you can install software. If you do not have such a computer, please let us know immediately so we can make alternative arrangements. You should bring your computer to class. During class we will have a "lab session" during which students will be experimenting and running code on their own computers.

Course Schedule

(subject to change)

- Week 1: Introduction to NLP and Basic Text Processing
- Week 2: Text Classification (Naive Bayes, MaxEnt, KNN, SVM, Decision Trees, Random Forests, etc.)
- Week 3: Feature extraction
- Week 4: Applications of Classifiers
- Week 5: Sentiment Analysis
- Week 6: Clustering Algorithms: k-means, dbscan, LDA
- Week 7: Factorization Algorithms: SVD, LSA, Word Embeddings
- Week 8: Applications of unsupervised algorithms
- Week 9: Advanced applications of NLP
- Week 10: Project Presentations