Data Science: Machine Learning and Natural Language Processing

---

Course Description:

Have you ever imagined how today's Hedge Funds process textual based data in real-time to make trading decisions? How it is that online ads seem to show content related to things you like? In addition to learning research methodologies, this course deals with inferring human behavior and sentiment through two core aspects of Data Science: Machine Learning and Natural Language Processing.

Goals:

1. To learn some basic research methodologies
2. To gain a background in the basics of python, machine learning and natural language processing
3. To work on a research project and write a paper for submission to a Journal

Prerequisites: Students expected to have a basic understanding of mathematics and basic familiarity with the Python Programming Language. Please have the following installed on your laptops prior to the start of the first day of class:

- Install Zotero add-on your laptops
    - This will help you track literature
- Install Python IDE – **version 3.0**
    - Anaconda distribution of Python

Individual Evaluations: Students will be evaluated based on the following criteria:

Class attendance consists of 10%
Project : 90%

A detailed grading rubric is below on next page of this syllabus:

		Category/Expectation	Meets (100%)	Below (75%)	Unsatisfactory (50% with submission 0% no submission)	Total Points
Attendance			Attended all classes and participated	Attended all classes but did not participate	Did not inform professor prior or failed to provide an explanation to as why a class was missed	10
Project	Presentation	Knowledge of Material	Presenters are fully prepared with understanding of the parameters.	Presenters are unprepared for the talk, demonstrating little understanding of how the topic relates to the defined parameters. .	N/A	10
		Content	Excellent visuals	Missing backup material where conclusions were inferred	N/A	10
	Research Paper	English Proficiency	High proficiency of English Language	Good Proficiency of English language	Minimal Proficiency of English language	30
		Content	An accurate and complete explanation of key concepts and theories is made. Enough detail is presented to allow the reader to understand the content and make judgments about it. Quality of paper is submission worthy to a journal.	The explanation is sufficiently inaccurate, incomplete, or confusing that the reader gains little information from the report. It appears that little attempt has been made to help the reader understand the material.	N/A	40

Homework collaboration policy: You may brainstorm and think through solutions with a small number of your classmates. However, you must write up your solutions entirely on your own. If you have used collaborators, you must state their names clearly next to your name on your write-up. Finally, copying solutions from the Internet or other textbooks is strictly prohibited.

Programming assignments: Each programming assignment MUST be completed and full reproducible using **Python 3.0**. You first need to install Python on your machine, if not already available. We recommend the Anaconda distribution:

- Download the anaconda distribution from http://continuum.io/downloads Anaconda includes and conveniently installs Python, the Jupyter Notebook, and other commonly used packages for scientific computing and data science.
- Please familiarize yourselves with this environment as soon as possible.

Course materials:

There is no specific textbook required. The professor will provide suggested books, papers and links for reference.

List of topics:

The topics covered will consist of the following:
Research Methodologies
Python
Machine Learning
Natural Language Processing

Research project:

You will work on a team based research project, example topics below. The project will count as 90% of your total grade, where you are required to provide a final presentation (you will present online during the last day of class) and research worthy paper for submission to a journal. Collaboration with your teams is required and your teammates will judge your participation in an end of semester anonymous review that will count towards your final grade. ** Kaggle is a great place to find data sets for your projects.

- Harnessing sentiment derived from Social Media (Twitter, StockTwits, etc.) to predict momentum effects in asset prices
- Create an Ideal Customer Profile using firmographic data
- Susceptibility to an ailment using anonymous medical data
- Probability of default using credit data
- Next day stock price prediction
- Likelihood of purchase based on online consumer behavior
- Movie recommendation

Detailed description of the course:

If it were not for the mass spend in R&D during the mid to late 1990s, we would still be hearing that crackling sound of the dial-up internet connection. Facebook and Google would be mere fractions of what they are today and Amazon would most likely be selling just books. Thankfully, through that massive investment in technology during the internet revolution, technological progression enabling the analysis of terabytes worth of data possible. From harnessing graphics processing units, GPUs, to perform deep learning to deploying entire architectures in the Amazon Cloud, both beautiful and amazing to say the least. You are amidst the biggest technological revolution your generation has ever experienced, hop aboard, and learn the skill-sets and get into that research-oriented mind-set to make a significant impact within the world of big data!

The goal of this class is to arm students with the necessary skill-set to perform quality research and to extract actionable insights from data and make predictions. This course will provide students with a foundational knowledge base of applied machine learning using the Python programming language. Students will learn vital data wrangling, feature selection, model selection and model validation techniques within statistical and probabilistic frameworks with an emphasis on *text analyti*cs and *natural language processing*. The intent is to not only expose students to modeling techniques but also place students in the mind-set to build a real working system through modules they create during both in-classroom and homework exercises. In addition, students will gain exposure to a variety of common tools used by Data Scientists by extracting insights and making predictions from AdTech, FinTech and MarTech datasets. Lastly, students will be in the position to submit their research paper to a journal for potential publication.

Course Schedule:

Following is the tentative course schedule:

Subject	Week	HW	Session	Topics
			1	Introduction to Research Methodologies, Journals and Submission Process
	Week 1		2	Data Wrangling, Cleaning Data, Dimension Reduction, Normalization, Imputation
		FORM TEAMS	3	Natural Language Processing: Text Tokenization, Stemming, Feature Matrix, Introduction
	Week 2		4	TF-IDF, Information Gain, Mutual Information, Feature Vector, N-gram methods, Embeddings, Literature Review
			5	Text summarization and Extraction, Topical modeling and key phrase extraction, Methodology
	Week 3	Research Questions Due	6	Sentiment Analysis: Lexicon and Machine Learning, Discussion, Model Selection, Grid Search, Validation and Evaluation, Performance Metrics
			7	Research Paper Review
ML and NLP	Week 4		8	Research Paper Review
			1	Individual Team Research Discussion
			2	Individual Team Research Discussion
		Introduction Due	3	Individual Team Research Discussion
			4	Individual Team Research Discussion
	Week 5	Data Section Due	5	Individual Team Research Discussion
			6	Individual Team Research Discussion
		Methodology Section Due	7	Individual Team Research Discussion
			8	Individual Team Research Discussion
		Results Section Due	9	Individual Team Research Discussion
Research	Week 6		10	Team Presentations