DS 200 Syllabus

Introduction to Data Sciences

Fall 2016

Tu Th 9:05 am - 10:20 pm, 210 IST

Instructor: Dr. John Yen

Office Hours: Th 2:30 - 3:30 pm, 313J IST

jyen@ist.psu.edu; (814) 865-6174

Teaching Assistant: Yafei Wang

Office Hours: Th 1:30 - 2:30 pm, 325 IST

yxw184@ist.psu.edu; (814) 441-5287

Goals of the Course: This course aims to achieve three goals.

1. First, it will provide you with hands-on experiences about a data science project, which will enable you to extract meaningful information (relevant to a question/hypothesis of interests to you) from a large twitter dataset you gather.
2. Second, you will learn four key concepts regarding predictive modeling and exploratory data analysis: Representation, Induction, Simplification, and Evaluation (RISE). This understanding will provide you a framework for relating theories (e.g., logic, probability) to practical methods using these theories (e.g., decision-tree induction, Naives Bayes induction), and their applications to data sciences, and to your data science project in particular.
3. Third, you will learn the broader landscape of Data Sciences.

- What global trends make Data Sciences important for our society?
- What are the "types" of data science projects and how, together, they form the journey of a data science initiative?
- What is the role of visual analytics in Data Sciences?
- What are the foundations of Data Sciences for innovating solutions for analyzing massive datasets?
- What should a Data Scientst know about data ethics?
- What is the role of domain-specific knowledge in Data Science projects?

While we may only be "touching the surface" of these topics, they will be addressed and elaborated in other courses throughout your Data Science education experience at Penn State. Together, I hope these goals help to guide you as you start this exciting journey of becoming the “Next Generation Data Scientists”.

Specific Learning Objectives:

Upon completion of the course, you should be able to gain first hand experiences about a mini data science project. More specifically, you will

- Be able to design an exploratory data science project using Tweets and assess its feasibility (using visualization tools).
- Be able to use Twitter API to gather tweets of interest for the project
- Be able to use a tool (Weka) to analyze twitter data
- Be able to use tools to visualize data and the models they generate
- Be able to generate a decision-tree predictive model for classifying tweets automatically using a tool (Weka)
- Be able to generate a probabilistic predictive model for classifying tweets automatically using a tool (Weka)
- Be able to evaluate and compare the performance of predictive models

You will be able to understand and apply the following concepts related to exploratory data analysis:

- R: Representation
- I: Induction
- S: Search
- E: Evaluation

You should also be able to gain a conceptual understanding about some of the real-world applications such as

- The "Beer and Diaper" data mining story-- The Discovery of Customer Purchase Patterns (Links to an external site.) (The discovery of frequent association (Links to an external site.); exploratory data project; human behavior; conditional probability)
- Amazon product recommendation based on reviews of others (Links to an external site.) (Similarity Measure; Collaborative Filtering; Recommendation Systems)
- Google's pre-processing of Web pages for Its Search Engine
- Social Media analytics
- Examples of data science applications in specific domains (e.g., health, social, security, life science).

Prerequisites: None

Organization of the Course

The class is designed such that each week consists of a lecture (typically on Tuesday) regarding theoretical, and/or technological foundations of data sciences, followed by a hands-on lab on Thursday. While all of these topics will be elaborated later through other Data Sciences courses, this course enables you to learn them in an integrated framework. Hopefully, you will find this

framework useful for you to incorporate additional knowledge and skills you learn from more focused and advanced data science course.

Data sciences are often described as the synergy created by three pillars: (1) computational pillar, (2) statistic pillar, and (3) domain expertise pillar. In this course, we organize topics into primarily three different modules: (1) four principles of exploratory data analysis: RISE, (2) two methods for predictive modeling, and (3) domain-agonistic topics of data sciences (e.g., theoretical foundations, data ethics, and the role of knowledge). Because topics in the first module are referred to in the second module, the second module can be viewed as an in-depth coverage of the four principles, with a focus on two types of induction methods.

Hands-on Labs and Mini Data Science Project

All hands-on labs are designed to help students to complete a mini data science term project. The project involves (1) using Twitter API to gather a set of tweets of interest, (2) filtering the irrelevant tweets using 2 predictive modeling methods (decision tree induction and Naives Bayes induction) and compare their performance, and (3) classifying the relevant tweets into two classes of interest (e.g., positive vs negative sentiment, support or against a presidential candidate, support or against a policy). Each hands-on lab introduces Python code examples, templates, libraries, or data analytics tools that will be useful for the project. Most of the hands-on labs on Thursday are related to the topics covered on Tuesday of the week. Unless noted otherwise, the hands-on labs are due 10 pm Sunday following the lab. Each project is completed by a team of 2 students. All project-related assignments are team assignments.

There are four deliverables related to the project:

- PD #1: Initial Project Idea and Feasibility: Describe the initial ideas regarding the project and initial feasibility assessment.
- PD #2: Feasibility Analysis of Project and Project Milestones: Describes initial data gathering and data tagging efforts that suggest the project is feasible. If the initial project idea is found to be infeasible, it needs to be revised and the feasibility of the revised idea need to be assessed. It is important that you document the iterative process of how ideas were created, tested for feasibility, and revised if needed until your team believe it is feasible. You will identify 3 milestones for completing the project. Each milestone is the goal of your lab from week 10 to week 12.
- PD #3: Filtering Tweet Report: Describe the result of filtering tweets using two models, and compare their performance.
- PD #4: Final Report: Describe the final results of classifying tweets relevant to your project goal using two models, compare their results, and discuss lessons learned.

Course Materials: Text Book: Doing Data Science by Cathy O'Neil & Rachel Schutt, O'Reilly.

Supplementary Reading materials will be posted in Canvas.

Week	Date	Topics	Lab	Due
Part 1		The RISE of Exploratory Data Analysis		

1	8/23,8/25	Introduction to Data Sciences The Story of Beer and Diaper; Why is Data Science important now (and in the future)?	Install Python; Initial Project Idea Brainstorming	PD #1: Initial Project Idea and Feasibility
2	8/30, 9/1	Exploratory Data Science Project (DSP) + Hypothesis-driven DSP = Data Science Initiative	Lab 2: Project-specific Twitter Data Gathering	Lab 2: Initial Project Tweets Gathering and Analysis
3	9/6, 9/8	The RISE of Predictive Modeling	Lab 3: Extended Project Tweets Gathering, and Tagging	Lab 3: Extended Project Tweets Gathering, Pre-processing, and Tagging Results
4	9/13, 9/15	Computational Representation of Text	Lab 4: Install Weka and pre-processing of Project Tweets	Lab 4: Install Weka and Pre-processing of Project Tweets
Par II		Predictive Modeling		
5	9/20, 9/22	Logic-based Model Representations, Decision Trees, and Information Gain	Lab 5 DT-based models using Weka	Lab 5 DT-based Models for Predicting Tweet Relevance
6	9/27, 9/29	Visual Analytics (Guest Lecturer: Dr. Luke Zhang)		
7	10/4, 10/6	Visual Analytics and Exploratory Data Analysis	Guest Lecture by Hamdan Azhar	PD #2: Project Milestones
	10/4	Lab 6: Temporal Visualization using Tableau		
8	10/11, 10/13	Bayesian Inference, Naïve Bayes Predictive Model Overfitting and Cross Validation	Lab 7: Naive Bayes Models and Cross-validation for Relevant Tweets	Lab 7 Cross-validation of Naïve Bayes Models for Relevant Tweets
9	10/18, 10/20	Spatial Data Analysis (Guest Lecturer: Dr. Jessie Li) Correlation, Questions, and Hypotheses	Lab 9	Project Deliverable PD3: Comparing DT and NB Models for Relevant Twee Classification
Par III		Domain-agnostic Topics of Data Sciences		
10	10/25, 10/27	Foundations for Data Sciences	Lab 10: Team MS #1	Lab 10: Team MS #1
11	11/1, 11/3	Data Ethics	Lab 11: Team MS #2	Lab 11: Team MS #2
12	11/8, 11/10	Domain-specific Knowledge and Data Sciences	Lab 12: Microsoft Band 2 Lab	Lab 12: Microsoft Band 2 Lab

13	11/15, 11/17	Guest Lecture: "Understanding people, places, and photos in social media" by Dr. Dhiraj Joshi, Senior Research Scientist, IBM T. J. Watson Research Using IBM Watson to make movie trailer (Links to an external site.)	Lab 13: Team MS #3	Lab 13: Team MS #3
	11/21-11/25	Thanksgiving Holiday		
14	11/29, 12/1	Project Report and Final Project Presentation (lab)	Final Project Presentations (student presentations)	
15	12/6, 12/8	Final Project Presentations (student presentations)	Final Project Presentations (student presentations)	
	12/12			PD #4: Final Project Report, Tagged Data, Model Outputs

Course Policies:

- Due to many in-class assignments of the course, attendance of the course is mandatory. Excused absences need to be approved by the instructor before the class to be missed. After three unexcused absences, penalty (10% of class attendance for each absence) will be applied to the final grade. A zero will be assigned (to the absent student) for each unexcused absence from in-class assignments.
- Late lab assignments and project-related assignments will receive a penalty of 25% for each day after the due date.
- Questions and class participation are encouraged and will be taken into consideration in the final grade.
- **Academic Integrity: According to the Penn State Principles and University Code of Conduct:** Academic integrity is the pursuit of scholarly activity in an open, honest and responsible manner. Academic integrity is a basic guiding principle for all academic activity at The Pennsylvania State University, and all members of the University community are expected to act in accordance with this principle. Consistent with this expectation, students should act with personal integrity, respect other students' dignity, rights and property, and help create and maintain an environment in which all can succeed through the fruits of their efforts. Academic integrity includes a commitment not to engage in or tolerate acts of falsification, misrepresentation or deception. Such acts of dishonesty violate the fundamental ethical principles of the University community and compromise the worth of work completed by others. Academic dishonesty includes, but is not limited to, cheating, plagiarism, fabrication of information or citations, facilitation of acts of academic dishonesty by others, unauthorized possession of examinations, submitting work of another person or work previously used without informing the

instructor, and tampering with the academic work of other students (also see Faculty Senate Policy 49-20 and G-9 Procedures).

- **Affirmative Action & Sexual Harassment:** The Pennsylvania State University is committed to a policy that all persons shall have equal access to programs, facilities, admission, and employment without regard to personal characteristics not related to ability, performance, or qualifications as determined by University policy or by Commonwealth or Federal authorities. Penn State does not discriminate against any person because of age, ancestry, color, disability or handicap, national origin, race, religious creed, sex, sexual orientation, or veteran status. Direct all inquiries to the Affirmative Action Office, 328 Boucke, University Park, PA 16802, (814) 863-0471.
- **Americans with Disabilities Act:** The College of Information Sciences and Technology welcomes persons with disabilities to all of its classes, programs, and events. If you need accommodations, or have questions about access to buildings where IST activities are held, please contact us in advance of your participation or visit. If you need assistance during a class, program, or event, please contact the member of our staff or faculty in charge.
- **An Invitation to Students with Learning Disabilities:** It is Penn State's policy to not discriminate against qualified students with documented disabilities in its educational programs. If you have a disability-related need for modifications in your testing or learning situation, your instructor should be notified during the first week of classes so that your needs can be accommodated. You will be asked to present documentation from the Office of Disability Services (located in 116 Boucke Building, 863-1807) that describes the nature of your disability and the recommended remedy. You may refer to the Nondiscrimination Policy in the Student Guide to University Policies and Rules.

Grading:

Evaluation of knowledge and understanding of materials will be based on works related to your class project (proposal, mid-term report, final project report, and project demo), lab assignments, and in-class engagements.

PD #1: Project Ideas and Initial Feasibility Assessment	5%
PD #2: Feasibility Analysis of Project and Milestones	5%
PD #3: Comparison of NB and DT for Filtering Tweets and Revised Milestones	10%
Term Project Final Presentation	10%
PD #4: Term Project Report	30%
In-class Engagement and Peer Evaluation	5%
Lab Assignments (not including PD's)	35%
Total	100%