INFO 6210
Data Management and Database Design
Spring 2020 Course Syllabus

Course Information

Professor: Nik Bear Brown
Email: nikbearbrown@gmail.com
Office: 505A Dana Hall
Office hours:
2:00 - 3:30 pm   F
12:30 - 2:00 pm M

Course website: Blackboard (for raw scores, uploading assignments, getting materials, & forums)

Piazza:

https:// piazza.com/northeastern/spring2020/info621008

Course Prerequisites

Engineering students only.

Course Description

Studies design of information systems from a data perspective for engineering and business applications; data modeling, including entity-relationship (E-R) and object approaches; user-centric information requirements and data sharing; fundamental concepts of database management systems (DBMS) and their applications; alternative data models, with emphasis on relational design; SQL; data normalization; data-driven application design for personal computer, server-based, enterprisewide, and Internet databases; and distributed data applications.

Communication

Communication between instructor and students is through

- Email via the Blackboard distribution list
- Announcements posted on Blackboard
- Notes posted on the Blackboard discussion board
- Private email exchanges

Course Structure

- Regularly test students on paper/algorithmic exercises
- Evaluate students' implementation competency, using assignments that require coding on given datasets
- Evaluate ability to setup data, code, and execute using python and SQL languages
- Student will be required to do "data digging": run analysis scripts and failure analysis
- Final project is typically asking and answering a "real world" question of interest using machine learning techniques

Learning Objectives

By the end of this course, students should be able to do the following:

Theory

Understand the Entity Relationship Model (ERM)
Understand Relational Algebra
Understand Relational Calculus

Conceptual knowledge

- Understand Relational Databases
- Understand SQL (SQL, NULL, integrity constraints, views, SQL Integrity constraints, outer join, SQL functions, user-defined aggregates, triggers)
- Understand SQL transactions
- Understand Storage and Indexing
- Understand Query Optimization
- Understand Normal Forms

Practical experience

- Collect clean and munge real-world data
- Store, search and display collected data.
- Display and search collected data via the web.

Course GitHub

The course GitHub (for all lectures, assignments and projects):

https://github.com/nikbearbrown/INFO_6210

nikbearbrown YouTube channel

Over the course of the semester I'll be making and putting additional data science and machine learning related videos on my YouTube channel.

https://www.youtube.com/user/nikbearbrown

The purpose of these videos is to put additional advanced content as well as supplemental content to provide additional coverage of the material in the course. Suggestions for topics for additional videos are always welcome.

Schedule

Note the first two weeks are a review of python and linux. It goes very fast because students are expected to have some basic programming abilities.

Week	Topic	Assignments
1) Week 1	**Intro Python ** Basic python ** Data Munging Python ** Data munging Data cleaning	Readings; Assignment 1
2) Week 2	** Theory ** Understand the Entity Relationship Model (ERM) Understand Relational Algebra Understand Relational Calculus	Readings HackerRank Online Quiz
3) Week 3	** Intro Linux on the Cloud ** Linux command line Setting up a server Database project one	Readings; Assignment 2
4) Week 4	** SQL ** Basic SQL Relational Database Model functions, user-defined aggregates, triggers	Readings; Project proposal HackerRank Online Quiz
5) Week 5	** SQL ** SQL transactions Indexing Query Optimization Normal Forms	Readings
6) Week 6	** SQL ** Distributed databases Backing up data	Readings; Assignment 3
7) Week 7	** SQL Example Database** Analyze hyperparameter database Project proposals	Readings HackerRank Online Quiz
8) Week 8	***No SQL** Intro No SQL Exam	Readings; Project progress report.
9) Week 9	***No SQL** MongoDB	Readings; Assignment 4 HackerRank Online Quiz

10) Week 10	**MapReduce/Hadoop** Introduction MapReduce/Hadoop	Readings HackerRank Online Quiz
11) Week 11	**MapReduce/Hadoop** MapReduce	Readings; Assignment 5
12) Week 12	Break	Spring recess
13) Week 13	**Graph Database** Graph Databases Neo4j	Readings; Assignment 6
14) Week 14	Research Project Presentations Exam	Readings; A draft of the final project for feedback

Teaching assistants

The Teaching assistants for Spring 2020 are:

Jayshil Jain <jain.ja@husky.neu.edu>
Newzy Sharma <sharma.new@husky.neu.edu>

Programming questions should first go to the TA's. If they can't answer them then the TA's will forward the questions to the Professor.

Learning Assessment

Achievement of learning outcomes will be assessed and graded through:

- Completion of assignments involving scripting in SQL and python, and analysis of data
- Completion of a term paper asking and answering a "real world" question of interest using machine learning techniques
- Portfolio piece
- Quizzes
- Exams

Reaching out for help

A student can always reach out for help to the Professor, Nik Bear Brown nikbearbrown@gmail.com. In an online course, it's important that a student reaches out early should he/she run into any issues.

Grading Policies

Students are evaluated based on their performance on assignments, performance on exams, and both the execution and presentation of a final project. If a particular grade is required in this class to satisfy any external criteria—including, but not limited to, employment opportunities, visa maintenance, scholarships, and financial aid—it is the student's responsibility to earn that grade by working consistently throughout the semester. Grades will not be changed based on student need, nor will extra credit opportunities be provided to an individual student without being made available to the entire class.

Grading Rubric

The following breakdown will be used for determining the final course grade:

Assignment	Percent of Total Grade
Assignments	40%
Quizzes	10%
Exams	25%
Projects	15%
Portfolio	10%

* Note that the assignments, presentations and drafts related to the projects go to that score rather than the programming assignments. I expect to use the following grading scale at the end of the semester. You should not expect a curve to be applied; but I reserve the right to use one.

Score	Grade
93 – 100	A
90 – 92	A-
88 – 89	B+
83 – 87	B
80 – 82	B-
78 – 79	C+
73 – 77	C
70 – 72	C-
60 – 69	D
<60	F

Scores in-between grades. For example, 82.5 or 92.3 will be decided based on the exams.

* Note the score is calculated using the grading rubric and IS NOT the average of the assignments that is displayed by BlackBoard.

Blackboard

You will submit your assignments via Blackboard *and* Github. Click the title of assignment (blackboard -> assignment -> <Title of Assignment>), to go to the submission page. You will know your score on an assignment, project or test via BlackBoard. BlackBoard only represents only the raw scores. Not normalized or curved grades. A jupyter notebook file ALONG with either a .DOC or .PDF rendering of that jupyter notebook file must be submitted with each assignment.

Multiple files must be zipped. No .RAR, .bz, .7z or other extensions.

Assignment file names MUST start with students last name then first name OR the groups name and include the class number and assignment number.

Assignment MUST estimate the percentage of code written by the student and that which came from external sources.

Assignment MUST specify a license at the bottom of each notebook turned in.

All code must adhere to a style guide and state which guide was used.

Due Dates

Due dates for assignments are usually every other Monday at midnight.

Five percent (i.e. 5%) is deducted for each day an assignment is late. Solutions will be posted the following Monday. Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date.

Course Materials

Required text (All free online)

Some textbooks are all available for free to NEU students via SpringerLink (http://link.Springer.com/). You must access SpringerLink from an NEU IP address to have full access and/or download these books.

If you are off-campus, in order to access resources provided through the Northeastern library outside the network, you should use their bookmarklet to load any page through the proxy: http://library.northeastern.edu/bookmarklet

Required Texts

The *required* textbooks we will be using in this class are:

Database Systems A Pragmatic Approach (2014) (This will be the primary book)
Authors: Elvis C. Foster, Shripad V. Godbole
ISBN: 978-1-4842-0878-6 (Print) 978-1-4842-0877-9 (Online)
http://link.springer.com/book/10.1007/978-1-4842-0877-9

Principles of Distributed Database Systems, Third Edition (2011)
Authors: M. Tamer Özsu, Patrick Valduriez
ISBN: 978-1-4419-8833-1 (Print) 978-1-4419-8834-8 (Online)
http://link.springer.com/book/10.1007/978-1-4419-8834-8

Beginning Database Design From Novice to Professional (2012)
Authors: Clare Churcher
ISBN: 978-1-4302-4209-3 (Print) 978-1-4302-4210-9 (Online)

http://link.springer.com/book/10.1007/978-1-4302-4210-9

Beginning Python
From Novice to Professional
Authors: Magnus Lie Hetland 2017
ISBN: 978-1-4842-0029-2 (Print) 978-1-4842-0028-5
https://link.Springer.com/book/10.1007/978-1-4842-0028-5

Beginning Django

Web Application Development and Deployment with Python
Daniel Rubio *(2017)*
https://link.springer.com/book/10.1007/978-1-4842-2787-9

Pro Django
Marty Alchin *(2013)*
https://link.springer.com/book/10.1007/978-1-4302-5810-0

Recommended Texts

Principles of Distributed Database Systems, Third Edition (2011)
Authors: M. Tamer Özsu, Patrick Valduriez
ISBN: 978-1-4419-8833-1 (Print) 978-1-4419-8834-8 (Online)
http://link.springer.com/book/10.1007/978-1-4419-8834-8

Beginning Database Design From Novice to Professional (2012)
Authors: Clare Churcher
ISBN: 978-1-4302-4209-3 (Print) 978-1-4302-4210-9 (Online)
http://link.springer.com/book/10.1007/978-1-4302-4210-9

The Definitive Guide to MongoDB: A complete guide to dealing with Big Data using MongoDB (2015)
Authors: David Hows, Peter Membrey, Eelco Plugge, Tim Hawkins
ISBN: 978-1-4842-1183-0 (Print) 978-1-4842-1182-3 (Online)
http://link.springer.com/book/10.1007/978-1-4842-1182-3

Pro Hadoop Data Analytics
Designing and Building Big Data Systems using the Hadoop Ecosystem
Authors: Kerry Koitzsch 2017
ISBN: 978-1-4842-1909-6 (Print) 978-1-4842-1910-2
https://link.springer.com/book/10.1007/978-1-4842-1910-2

Pro Apache Hadoop
Authors: Sameer Wadkar, Madhu Siddalingaiah 2014
ISBN: 978-1-4302-4863-7 (Print) 978-1-4302-4864-4

https://link.springer.com/book/10.1007/978-1-4302-4864-4

Pro Spark Streaming
The Zen of Real-Time Analytics Using Apache Spark
Authors: Zubair Nabi 2016
ISBN: 978-1-4842-1480-0 (Print) 978-1-4842-1479-4
https://link.springer.com/book/10.1007/978-1-4842-1479-4

Pro Python Best Practices
Debugging, Testing and Maintenance
Authors: Kristian Rother 2017
ISBN: 978-1-4842-2240-9 (Print) 978-1-4842-2241-6 (Online)
https://link.Springer.com/book/10.1007/978-1-4842-2241-6

Python Recipes Handbook
A Problem-Solution Approach
Authors: Joey Bernard 2016
ISBN: 978-1-4842-0242-5 (Print) 978-1-4842-0241-8
https://link.Springer.com/book/10.1007/978-1-4842-0241-8

Lean Python
Learn Just Enough Python to Build Useful Tools
Authors: Paul Gerrard 2016
ISBN: 978-1-4842-2384-0 (Print) 978-1-4842-2385-7
https://link.Springer.com/book/10.1007/978-1-4842-2385-7

Learn to Program with Python
Authors: Irv Kalb 2016
ISBN: 978-1-4842-1868-6 (Print) 978-1-4842-2172-3
https://link.Springer.com/book/10.1007/978-1-4842-2172-3

Big Data Made Easy
A Working Guide to the Complete Hadoop Toolset
Authors: Michael Frampton 2015
ISBN: 978-1-4842-0095-7 (Print) 978-1-4842-0094-0
https://link.springer.com/book/10.1007/978-1-4842-0094-0

The Definitive Guide to SQLite (2010)
Authors: Grant Allen, Mike Owens
ISBN: 978-1-4302-3225-4 (Print) 978-1-4302-3226-1 (Online)
http://link.springer.com/book/10.1007/978-1-4302-3226-1

The Definitive Guide to MongoDB: A complete guide to dealing with Big Data using MongoDB (2015)
Authors: David Hows, Peter Membrey, Eelco Plugge, Tim Hawkins
ISBN: 978-1-4842-1183-0 (Print) 978-1-4842-1182-3 (Online)
http://link.springer.com/book/10.1007/978-1-4842-1182-3

Beginning CouchDB (2009)
Authors: Joe Lennon
ISBN: 978-1-4302-7237-3 (Print) 978-1-4302-7236-6 (Online)
http://link.springer.com/book/10.1007/978-1-4302-7236-6

Software

python Anaconda

- https://www.continuum.io/anaconda-overview

R (Statisical programming language)

- R project https://www.r-project.org/

RStudio (IDE)

- RStudio https://www.rstudio.com/products/rstudio/download3/

Python Tutorials

Dive into Python http://diveintopython.org

Python 101 – Beginning Python http://www.rexx.com/~dkuhlman/python_101/python_101.html

The Official Python Tutorial http://www.python.org/doc/current/tut/tut.html

The Python Quick Reference http://rgruet.free.fr/PQR2.3.html

Python Fundamentals Training – Classes http://www.youtube.com/watch?v=rKzZEtxIX14

Python 2.7 Tutorial Derek Banas· http://www.youtube.com/watch?v=UQi-L-_chcc

Python Programming Tutorial - thenewboston http://www.youtube.com/watch?v=4Mf0h3HphEA

Google Python Class http://www.youtube.com/watch?v=tKTZoB2Vjuk

Nice free CS/python book https://www.cs.hmc.edu/csforall/index.html

datacamp.com https://www.datacamp.com/tracks/python-developer

Participation Policy

Participation in discussions is an important aspect on the class. It is important that both students and instructional staff help foster an environment in which students feel safe asking questions, posing their opinions, and sharing their work for critique. If at any time you feel this environment is being threatened—by other students, the TA, or the professor—speak up and make your concerns heard. If you feel uncomfortable broaching this topic with the professor, you should feel free to voice your concerns to the Dean's office.

Collaboration Policies

Students are strongly encouraged to collaborate through discussing strategies for completing assignments, talking about the readings before class, and studying for the exams. However, all work that you turn in to me with your name on it must be in your own words or coded in your own style. Directly copied code or text from any other source MUST be cited. In any case, you must write up your solutions, in your own words. Furthermore, if you did collaborate on any problem, you must clearly list all of the collaborators in your submission. Handing in the same work for more than one course without explicit permission is forbidden.

Feel free to discuss general strategies, but any written work or code should be your own, in your own words/style. If you have collaborated on ideas leading up to the final solution, give each other credit on what you turn in, clearly labeling who contributed what ideas. Individuals should be able to explain the function of every aspect of group-produced work. Not understanding what plagiarism is does not constitute an excuse for committing it. You should familiarize yourself with the University's policies on academic dishonesty at the beginning of the semester. If you have any doubts whatsoever about whether you are breaking the rules – ask!

Any submitted work violating the collaboration policies WILL BE GIVEN A ZERO even if "by mistake." Multiple mistakes *will be sent to OSCCR for disciplinary review.*

To reiterate: **plagiarism and cheating are strictly forbidden. No excuses, no exceptions.** *All incidents of plagiarism and cheating will be sent to OSCCR for disciplinary review.*

Assignment Late Policy

Assignments are due by 11:59pm on the due date marked on the schedule. Late assignments will receive a 5% deduction per day that they are late, including weekend days. It is your responsibility to determine whether or not it is worth spending the extra time on an assignment vs. turning in incomplete work for partial credit without penalty. Any exceptions to this policy (e.g. long-term illness or family emergencies) must be approved by the professor.

Five percent (i.e. 5%) is deducted for each day an assignment is late. Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date.

Only ONE extension will be granted per semester.

Student Resources

Special Accommodations/ADA: In accordance with the Americans with Disabilities Act (ADA 1990), Northeastern University seeks to provide equal access to its programs, services, and activities. If you will need accommodations in this class, please contact the Disability Resource Center (www.northeastern.edu/drc/) *as soon as possible* to make appropriate arrangements, and please provide the course instructors with any necessary documentation. The University requires that you provide documentation of your disabilities to the DRC so that they may identify what accommodations are required, and arrange with the instructor to provide those on your behalf, as needed.

Academic Integrity: All students must adhere to the university's Academic Integrity Policy, which can be found on the website of the Office of Student Conduct and Conflict Resolution (OSCCR), at http://www.northeastern.edu/osccr/academicintegrity/index.html. Please be particularly aware of the policy regarding plagiarism. As you probably know, plagiarism involves *representing anyone else's words or ideas as your own*. It doesn't matter where you got these ideas—from a book, on the web, from a fellow-student, from your mother. It doesn't matter whether you quote the source directly or paraphrase it; if you are not the originator of the words or ideas, *you must state clearly and specifically where they came from*. Please consult an instructor if you have any confusion or concerns when preparing any of the assignments so that together. You can also consult the guide "Avoiding Plagiarism" on the NU Library Website at http://www.lib.neu.edu/online_research/help/avoiding_plagiarism/. If an academic integrity concern arises, one of the instructors will speak with you about it; if the discussion does not resolve the concern, we will refer the matter to OSCCR.

Writing Center: The Northeastern University Writing Center, housed in the Department of English within the College of Social Sciences and Humanities, is open to any member of the Northeastern community and exists to help any level writer, from any academic discipline, become a better writer. You can book face-to-face, online, or same day appointments in two locations: 412 Holmes Hall and 136 Snell Library (behind Argo Tea). For more information or to book an appointment, please visit http://www.northeastern.edu/writingcenter/.