BUDT 758-0501/2/3: Big Data and Artificial Intelligence for Business (Fall 2020)

Decisions, Operations & Information Technologies
Robert H. Smith School of Business

Instructor: Kunpeng Zhang (kpzhang@umd.edu)
Lecture-Discussions: Tuesday/Thursday, 2:00 – 3:15 (0501), 3:30 – 4:45 (0502), 5:00 – 6:15 PM (0503)
Room: **Online**
Zoom link: https://umd.zoom.us/j/3206804434

Office Hour: TBD
TA: TBD

Textbook: Mining of Massive Datasets
Hardcopy: Amazon.com
E-version: Free available: http://infolab.stanford.edu/~ullman/mmds/book.pdf

Deep Learning
E-version: http://www.deeplearningbook.org/

About the course

Big data represents unprecedented opportunities for companies to generate insights and create wealth. Internet of Things (IoT) is connecting almost all the components together in every aspect of business and our daily life. As a result, huge amount of data is being generated. At the same time, much of the big data is unstructured, in real time and only loosely connected. It defies the traditional ways of managing databases. This creates challenges even to tech-savvy companies on how to leverage the big data to gain competitive advantages. Challenges and opportunities coexist. To extract the great value from the data, we should be equipped with advanced techniques. Artificial Intelligence (AI) is penetrating our daily routines deeply, and shows great promise in exciting areas such as healthcare and autonomous driving cars.

This course uses a hands-on, learning-by-doing approach to understanding the concepts behind Big Data, IoT, and AI, the strategic drivers of these technologies and the value propositions that they provide to industries. In addition, the course will also serve as an introduction to some of the key technologies within this ecosystem, such as Hadoop, AWS, Pig, Hive, Amazon Web Services and Spark. Examples of AI using Deep Learning will be conducted in class. The focus is on creating awareness of the technologies, allowing some level of familiarity with them through assignments, and enabling some strategic thinking around the use of these in business.

The technologies are still evolving very rapidly. Therefore, there is a level of experimentation with new material that will take place during the semester. Students are required to be flexible as and when topics or material in class are revised or modified. We will do our best to ensure that no undue burden is placed on students.

Learning Objectives

The course has two primary objectives:

1. To allow students to have working knowledge and exposure to key elements of a big data technology platform, and a basic AI example
2. To allow students to understand critical business and strategic issues around the use of these technologies in organizations and to help guide the successful design and implementation of complex data strategy

Though mastery of this content requires more than one course, an introductory course, such as this, is useful in allowing students to gain much-needed familiarity with these technologies and concepts. That is the objective of this course.

Prerequisites

1. Python programming
2. Databases, specifically working knowledge of relational databases and SQL
3. Working knowledge of Linux/Unix is useful but not required.

Software Needed

Much of the software needed for big data applications tend to be open source. Therefore, the source programs are free and will be provided in class. We will be using Amazon Web Services (AWS), Google Colab, as well as Pytorch and Keras.

1. Amazon Web Services – provided by instructors (Hadoop, Pig, Hive, Impala)
2. Python + Torch / Keras – open source

Required Reading Material

A significant proportion of the reading material for this course is available online and is free. When necessary, additional reading material will be posted on Canvas/ELMS.
Optional useful sources are listed below; these are not required but are good reference material.

1. Hadoop: The Definitive Guide, by Tom White (http://it-ebooks.info/book/5629/)
2. Big Data: A Revolution That Will Transform How We Live, Work, and Think, by Viktor MayerSchonberger and Kenneth Cukier (http://www.big-data-book.com/)
3. Deep Learning, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville (https://github.com/HFTrader/DeepLearningBook/blob/master/DeepLearningBook.pdf)

Course Format and Grading

Classes

We **virtually** meet twice each week. Tuesdays will be mostly lecturing, and Thursdays will be mostly in-class exercises.

Assignments/Quiz

We have lab assignments, Quiz, and individual reports throughout the semester.

Class project

There is a class project for each group. The size of each group is **3** at maximum, and students will be assigned to groups by the instructors. Each team will find a dataset and a specific set of questions associated with that dataset – the team will be responsible for analyzing the data and providing a series of analyses that will help answer the questions. In addition, the team will provide a short report on the project outcomes, including screen shots and samples of code, as specified. Each will also have to participate the poster slam competition. Further details of the projects will be provided in class.

Grading

Your final grade for the course will be composed from the following items:

Class participation: 10%*1 =10%
Lab Assignment: 5%*6 = 30%
Quiz: 10% + 10% + 10% = 30%
Class project: 30%*1 = 30%

Attendance

I assume that you understand the importance of attending class. While I do not check attendance in every lecture, I expect you to be present unless circumstances make that impossible. If you miss **TWO lectures**, you will definitely **NOT receive A** for your final grade.

If you miss your project presentation or Quiz without an extremely good excuse, you will receive a grade of **ZERO** for that. If you think you have an excuse for missing your presentation or Quiz, please discuss it with me, **in advance if possible**. If I judge that your excuse is reasonable, I will -- depending on the circumstances -- either give you a make-up presentation, or I will average your other grades so that the missing grade does not count against you.

Although it should not need to be said, I expect you to maintain a reasonable level of decorum in class. This means that there is usually no eating or drinking in class. Cell phones are suggested to be turned off. You'd better not walk in late or walk in and out of the room during lecture.

Disability Services

The Office of Disability Services works to ensure the accessibility of UMD programs, classes, and services to students with disabilities. Services are available for students who have documented disabilities, including vision or hearing impairments and emotional or physical disabilities. Students with disability/access needs or questions may contact the Office of Disability Services at (301) 314-7682.

Academic Integrity

The Robert H. Smith School of Business recognizes honesty and integrity as necessary cornerstones to the pursuit of excellence in academic and professional business activities. The University's Code of Academic Integrity is designed to ensure that the principles of academic honesty and integrity are upheld. All students are expected to adhere to this Code. The Smith School does not tolerate academic dishonesty. All acts of academic dishonesty will be dealt with in accordance with the provisions of this code. Please visit the following website for more information on the University's Code of Academic Integrity: http://www.inform.umd.edu/CampusInfo/Departments/JPO/AcInteg/code_acinteg2a.html

Plagiarism Policy: Inevitably in a programming course, it seems that a few people will turn in work that is not their own. You should understand that it is usually easy to detect copying of programs -- even when a program is modified to try to disguise its source. Copying a program, or letting someone else copy your program, is a form of academic dishonesty and the penalties can be found **here**.

Tentative Schedule

Here is a tentative schedule of lectures, readings, and labs for this course. We will try to keep approximately to this schedule. We will not cover every topic in every section -- but I recommend you to read the first seven chapters of the book in their entirety, if you are really interested in learning Java.

(Note that we may change the schedule during the semester.)

Topics	Lab	Assignment
1. Introduction - Business Value of Big data and AI - Internet of Things - Review of machine learning		

2. Deep Learning (1) - Introduction - Fully Connected Feedforward		
	Lab 1: Sentiment prediction using deep neural networks (dataset: doctor reviews)	Submit report
3. Deep Learning (2) - Batch normalization - Auto-encoder - Word embedding - Topic modeling		
	Lab 2: word2vec	Submit report
4. Deep Learning (3) - CNN - RNN (LSTM)		
	Lab 3: Sentiment prediction using CNN and RNN	Submit report
5. Deep Learning (4) - Generative Adersarial Network		
	Handwritten digit generation (practice)	
		Quiz 1
6. Cloud computing - AWS		**Project proposal Due**
7. Hadoop - Overview of Hadoop Ecosystem - HDFS		
8. MapReduce		
9. Data Management - RedShift - MongoDB - Hive		
10. Data Pipelines - Glue - Athena - Sage maker		
	Lab 4 - RedShift, Hive - Glue, Athena - Sage maker	Submit report
		Quiz 2
11. Spark - Introduction - RDD - DataFrame		
	Lab 5: Data operations using Spark	
12. Spark - SQL - ML/GraphX		
	Lab 6: Clustering and network analysis using Spark	

		Quiz 3
		Project report due