Syllabus for DSBA 6345 – Modern Data Science Systems
(The topics of the class are
(Cross listed as ITCS 6345)

Note (Jan/2021): This is a somewhat updated but previously used syllabus. The sources and reference have been updated (but additional updates will be necessary). The new topics one will have more big-data/cloud-related stuff. Also, in the online edition, we’ll need to focus more on projects.

Projects. Your grade will depend on your group performance against everybody else participating in the competition; or if the project is a research project, on the publication of your article. This will be challenging, but will also give you an opportunity to shine. My former students cashed on their good results by securing jobs with leading companies.

• Look at the serious competitions on Kaggle.com and CodaLab.org
• Look into BioAsq,
• See if there are any topics you’re interested in.
• The sweet spot for many competitions is when the ‘accuracy’ is above random but below 90/%. Focus your attention on these. Pay attention to the dates and rules – we’ll have to obey them.
• I also have a couple of small research projects; these are technically not very difficult (github + python notebooks) but require reading and writing beyond what is happening in class.

Finally, the total number of units in the outline below exceeds our available time, and therefore we’ll need to focus on some topics, and disregard other themes. Think about which topics we should cover. We’ll definitely spend lots of time on deep learning (attention); topological data analysis (TDA) will also be covered.

See you in class,
WZ

Instructor: Dr. Wlodek Zadrozny, Department of Computer Science, College of Computing and Informatics

Office: 435D Woodward (online, by appointment)

Email: wzadrozn@uncc.edu

Course Description: This class will focus on more advanced and recent techniques in data science, and their applications to business problems. The topics will include enterprise search and question answering, machine learning with neural networks, probabilistic and graph algorithms, and topological data analysis. Most of the assignments will be done ‘in the cloud’ and on Google Colab. The class will assume basic knowledge of data science exemplified by a graduate, undergraduate or online machine learning class; a natural language processing class; as well as knowledge of elementary linear algebra, calculus and statistics. Some experience with cloud computing is also expected.

Credit hours: 3
Pre-requisites: DSBA/ITCS 6100

Objectives of the course:
- Learn about advanced and recent techniques in data science and their applications to business problems
- Understand enterprise search and question answering
- Understand deep neural networks, probabilistic and graph algorithms, and topological data analysis

Instructional Method: Lecture + Student Presentations

Means of student evaluation (tentative):
Two 45-60 min Midterms – 30%
Homework and class participation -- 15%
Projects – 35 %
Final exam 20%

Grading:
90% and above A (Superior Performance)
80% - 89.9% B (Good Performance)
65% - 79.9% C (Average Performance)
Below 65% U (Unsatisfactory)

Other policies that apply to this course:

Attendance:
Students are expected to attend all class meetings and to arrive before the class starts. Class topics are integrated, with each week building on prior weeks. Failure to attend or to arrive on time can adversely affect both individual performance, ability to contribute to the group project, and the earned letter grade. If a student misses 4 weeks of class or more, they will automatically receive an unsatisfactory U grade in the course regardless of earned points to date on other activities. If a student misses a class due to work or other reasons, it is their responsibility to get notes from peers; instructors do not hold extra repeat class sessions.

Academic integrity:
Students have the responsibility to know and observe the requirements of The UNC Charlotte Code of Student Academic Integrity available online at [http://legal.uncc.edu/policies/up-407](http://legal.uncc.edu/policies/up-407). This code forbids cheating, fabrication or falsification of information, multiple submissions of academic work, plagiarism (which includes viewing others work without instructor permission), abuse of academic materials, and complicity in academic dishonesty. This forbidding includes sharing/copying work between individuals or teams without permission of instructors. Any special requirements or permission regarding academic integrity in this course will be stated by the instructor, and are binding on the students. Students who violate the code can be expelled from UNC Charlotte. The normal penalty for a first offense is zero credit on the work involving dishonesty and further substantial reduction of the course grade. In almost all cases the course grade is reduced to failing. Students are expected to report cases of academic dishonesty to the course instructor.

Inclement Weather:
University Policy Statement #13 states the University is open unless the Chancellor announces that the University is closed. The inclement weather hotline number to call is 704-786-2877. In the event of inclement weather, check your email the morning of class. The instructors will use their best judgment as to whether class should be held understanding that some of you commute from far away and the instructors will notify you by email if class is cancelled.
Textbooks or resources:

- [MMDS] Mining massive data sets: http://www.mmds.org/
  - https://online.stanford.edu/courses/soe-ycs0007-mining-massive-data-sets (requires registration)
  - https://www.youtube.com/channel/UC_Oao2FYkLAUlUVkBFze4ig/videos (older version on youtube)
  - Social and Information Network Analysis http://web.stanford.edu/class/cs224w/
  - Project in Mining Massive Data Sets http://web.stanford.edu/class/cs341/info.html

Additional resources:

- https://github.com/aditya-grover/node2vec/tree/master/src
- ...

Software (all resources listed below are available for free to students; students are expected to bring their own laptops to class):

1. Apache Lucene/Solr Laptop and Cloud
2. Tensorflow, Word2vec, Glove, Gensim (or equivalent)
3. Ripser (or related, e.g. Mapper)
4. SNAP http://snap.stanford.edu/proj/snap-www (or equivalent)

Cloud. One or more of the following:

1. Google Cloud and/or Amazon AWS
2. Microsoft Azure
3. IBM?

Topical outline of course content
<table>
<thead>
<tr>
<th>Each unit = 1 week of classes unless noted otherwise</th>
<th>1. Introduction (1-2 Units)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Lecture: Introduction to advanced data mining with the focus on search (Vector-space model for text representation, and Boolean information retrieval; Similarity measures )</td>
</tr>
<tr>
<td></td>
<td>• Hands-on: Explore Lucene index with Luke; Install Lucene or Solr; Create an index. (Local and cloud).</td>
</tr>
<tr>
<td></td>
<td>• Readings: IR-book: Ch1-3; 19-20; MMDS Ch 1.</td>
</tr>
<tr>
<td>2. Information retrieval models (1-2 units)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Lecture: Models of search: Page Rank, HITS; Boolean, BM25</td>
</tr>
<tr>
<td></td>
<td>• Hands-on: Matrix operations in Python on term-document matrices; PageRank convergence on citations data and Enron data;</td>
</tr>
<tr>
<td></td>
<td>• Readings:</td>
</tr>
<tr>
<td></td>
<td>o [IR-book]: Ch 4-6 and Ch 21;</td>
</tr>
<tr>
<td></td>
<td>o [MMDS] Ch5.</td>
</tr>
<tr>
<td>3. Question answering, dialogue systems and chatbots (2 units)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Lectures: IBM Watson; Question answering using deep learning</td>
</tr>
<tr>
<td></td>
<td>• Hands-on: answering factoid questions using Wikipedia search; improving results using ensemble learning and internet search.</td>
</tr>
<tr>
<td></td>
<td>• Readings: AAAI Mag. 2011.; TBD.</td>
</tr>
<tr>
<td>4. Dimensionality reduction techniques (2 units)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Lectures: Matrix decomposition, Iterative methods, Random projections/LSH, SVD</td>
</tr>
<tr>
<td></td>
<td>• Hands-on: Python NumPy, SciPy and Matplotlib</td>
</tr>
<tr>
<td></td>
<td>• Readings:</td>
</tr>
<tr>
<td></td>
<td>o [MMDS] Ch11.</td>
</tr>
<tr>
<td>5. Deep learning and distributional semantics (3-4 units)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Lectures: Why deep learning? Neural network on text: word2vec, glove and related models</td>
</tr>
<tr>
<td></td>
<td>• Hands-on: Tensorflow, gensim, word2vec, glove (one or more of these)</td>
</tr>
<tr>
<td></td>
<td>• <strong>Attention-based models:</strong> BERT and similar (new)</td>
</tr>
<tr>
<td></td>
<td>• Readings:</td>
</tr>
<tr>
<td></td>
<td>o [JM] Ch16 Semantics with Dense Vectors</td>
</tr>
<tr>
<td></td>
<td>o [DL] selected sections of Part II or equivalent</td>
</tr>
<tr>
<td>6. Probabilistic and graph algorithms on social network data (2-3 units)</td>
<td></td>
</tr>
</tbody>
</table>
• Lecture: Mining Social-Network (and other) Graphs
• Hands-on: http://snap.stanford.edu
• Readings: [MMDS] Ch10

7. Topological data analysis (2-4 units)
• Lecture: Basic concepts of topological data analysis: Data has shape and shape matters.
• Hands-on: Ripser (or similar tool in Python, e.g. Mapper)
• Readings: Barcodes: The persistent topology of data. : Robert Ghrist
  Journal: Bull. Amer. Math. Soc. 45 (2008), 61-75

• Video: The Shape of Data, by Gunnar Carlsson
  https://www.youtube.com/watch?v=iOxLgbnl1u4
  https://www.youtube.com/watch?v=X9ktWgJ7ung