CS795/895: Mining Scholarly Big Data Syllabus (Fall 2020)

Instructor

Jian Wu

Email

jwu@cs.odu.edu

Office Location

Zoom

Office Hours

11-noon Monday or by appointment

Class Time

9:30 am. -10:45 am. M/W

Important Dates

Monday, 8/31/2020: first class

Tuesday, 9/8/2020: Add/Drop deadline

Wednesday 12/9/2020: last class

Course Overview

One of the computer science subject areas that was impacted by artificial intelligence in the last decade is text mining. The subject covers many extensively studied topics in scholarly big data that drew interest from both academia and industry researchers. This seminar is designed for graduate students to explore important research topics in mining scholarly big data such as searching, metadata extraction, citation parsing, figure/table extraction, entity and relation extraction, scientific trend analysis (science-of-science), recommender systems, document representation learning, and question answering systems. In the first half of the class, students will learn fundamentals of NLP, machine learning and deep learning, data mining and big data. The other half of the seminar will cover contemporary research topics in scholarly big data. Students will have opportunities to have hands-on practice on a Spark cluster to process scholarly big data using machine learning techniques to replicate or research recently published works in prestigious conferences.

Course Delivery Method

Due to the health concerns amid the COVID-19 pandemic, this course will be held on ZOOM sessions for the whole semester. Slides will be available. Recording will not be available. Connection instructions will be sent to students before class begins.

Required Text

There is no required textbook. Two recommended textbooks are

- ○ Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman et al. 2020
- ○ Learning Spark: Lightning-Fast Data Analysis 2nd Edition, Jules S. Damji, Brooke Wenig, 2020

Hardware and Software Requirements

Students will need frequent access to a PC (with Windows 10) or a Mac (with MacOS 10.14+) capable of hosting application development activities or of connecting to remote servers. Students will be attending network conferences requiring the use of a microphone. Webcams are optional. For both remote access to servers and for network conferencing, a good-quality internet connection is important.

Students will have Zoom installed on their computers. The course will introduce students to a wide variety of open-source, free software, but students will need to install some of these on their chosen development machine.

Course Materials

- Course materials and other resources including slides and assignments will be distributed as the course proceeds in the semester.

Grading Policy

Students are graded based on the following aspects.

- Attendance: 10%
- Pop quizzes: 10% (5, each 2%)
- Warm up project: 10%
- Project: 50% (evaluated based on presentation and reports)
    - Proposal: 10%, Milestone 1: 10%, Milestone 2: 10%, Final: 20%
- Student presentation: 20%

Grading Chart

A	A-	B+	B	B-	C+	C*
94-100	90-93.99	87-89.99	84-86.99	80-83.99	77-79.99	74-76.99

* A provisional graduate student who receives one C in any of the required prerequisites will be subject to removal from the graduate program. A graduate student must maintain at least a 3.0 grade point average to graduate. (ODU Grading Policy)

Attendance Policy

Attendance is required. One absence causes a deduction of 1% on attendance until all points are deducted in this aspect. If more than 11 absences are observed, the student automatically get F for this course. In case of absence due to legitimate reasons, including but not limited to sickness, University-approved curricular and extracurricular activities (such as athletic contests), career interviews, the death of family members, students should be prepared to provide documentation **before** classes. Makeup classes are not available, but students can always discuss with the instructor about course content in office hours.

Academic Integrity

Individual assignments must be completed independently. Students are strongly encouraged to form study groups and to learn from their peers. However, discussion on final proposal writing and presentation in the study group should be limited to general approaches to solutions. **Specific answers should never be discussed**. ODU's policy regarding Academic Integrity must be followed.

- **Cheating**: Using unauthorized assistance, materials, study aids, or other information in any academic exercise (Examples of cheating include, but are not limited to, the following: using unapproved resources or assistance to complete an assignment, paper, project, quiz or exam; collaborating in violation of a faculty member's instructions; and submitting the same, or substantially the same, paper to more than one course for academic credit without first obtaining the approval of faculty).
- **Plagiarism**: Using someone else's language, ideas, or other original material without acknowledging its source in any academic exercise. 4 Examples of plagiarism include, but are not limited to submitting a research paper obtained from a commercial research service, the Internet, or from another student as if it were original work; or making simple changes to borrowed materials while leaving the organization, content, or phraseology intact. Plagiarism also occurs in a group project if one or more of the members of the group does none of the group's work and participates in none of the group's activities but attempts to take credit for the work of the group.
- **Fabrication**: Inventing, altering or falsifying any data, citation or information in any academic exercise. Examples of fabrication include, but are not limited to, the following: citation of a primary source which the student actually obtained from a secondary source; or invention or alteration of experimental data without appropriate documentation (such as statistical outliers).
- **Facilitation**: Helping another student commit, or attempt to commit, any Academic Integrity violation, or failure to report suspected Academic Integrity violations to a faculty member. An example of facilitation may include circulating course materials when the faculty member has not explicitly authorized their use.

Copyright

- All course materials students receive or to which students have online access are protected by copyright. Students may use course materials and make copies for their own use as needed, but **unauthorized distribution and/or uploading of materials without the instructor's express permission is strictly prohibited**.

Disability Accommodations

- In order to receive consideration for reasonable accommodations, you must contact the appropriate services office will provide you with an accommodation letter. Please share this letter with your instructors and discuss the accommodations with them as early in your courses as possible. The detail of disability accommodations is documented in ODU policy #4500.

Discrimination and Harassment

- The university is committed to equal access to programs, facilities, admission and employment for all persons. It is the policy of the university to maintain an environment free of harassment and free of discrimination against any person because of age, race, color, ancestry, national origin, religion, creed, service in the uniformed services (as defined in state and federal law), veteran status, sex, sexual orientation, marital or family status, pregnancy, pregnancy-related conditions, physical or mental disability, gender, perceived gender, gender identity, genetic information or political ideas. Discriminatory conduct and harassment, as well as sexual misconduct and relationship violence, violates the dignity of individuals, impedes the realization of the university's educational mission, and will not be tolerated.
- Gender-based sexual harassment, including sexual violence, are forms of gender discrimination in that they deny or limit an individual's ability to participate in or benefit from University programs or activities. These policies shall not be construed to restrict academic freedom at the university, nor shall they be construed to restrict constitutionally protected expression. The policy is coded in University Policy #1005.

Course Schedule*

Week	Dates	Subject	Practice Problems
1	Monday, 8/31/2020	Course Introduction	
1	Wednesday, 9/2/2020	An Introduction to Scholarly Big Data and Spark cluster logging in	Logging in and testing Spark cluster due
2	Monday, 9/7/2020	Labor Day Holiday (no classes)	

Week	Dates	Subject	Practice Problems
2	Wednesday, 9/9/2020	Warmup project and Fundamentals of NLP 1	Spark cluster warm up project
3	Monday, 9/14/2020	Contemporary research problems in Mining Scholarly Big Data	Preliminary project selection
3	Wednesday, 9/16/2020	Fundamentals of NLP 1	
4	Monday, 9/21/2020	Fundamentals of NLP 2	Quiz 1
4	Wednesday, 9/23/2020	Fundamentals of NLP 3	Quiz 2
5	Monday, 9/28/2020	Project proposal presentation	Project proposal due
5	Wednesday, 9/30/2020	Fundamentals of Data Mining 1	Student presentation assignments
6	Monday, 10/5/2020	Fundamentals of Data Mining 2	Warm up project report due; Quiz 3
6	Wednesday, 10/7/2020	Fundamentals of Data Mining 3	Quiz 4
7	Monday, 10/12/2020	Fundamentals of Data Mining 4	Quiz 5
7	Wednesday, 10/14/2020	Milestone 1 presentation	Milestone 1 report due
8	Monday, 10/19/2020	Student presentation: Text Summarization (Muntabir)	
8	Wednesday, 10/21/2020	CiteSeerX: AI in Scholarly Big Data	
9	Monday, 10/26/2020	Student presentation: Science of Science (Erica)	
9	Wednesday, 10/28/2020	Invited talk (Juan Castorena, LANL)	
10	Monday, 11/2/2020	Student presentation: Content-based Image Search (Martin)	
10	Wednesday, 11/4/2020	Milestone 2 presentation	Milestone 2 report due

Week	Dates	Subject	Practice Problems
11	Monday, 11/9/2020	Student presentation: Research Paper Recommender Systems (Richard)	
11	Wednesday, 11/11/2020	Invited talk (Fengjiao Wang, ODU)	
12	Monday, 11/16/2020	Student presentation: Document Classification (Xin)	
12	Wednesday, 11/18/2020	Invited talk (Lu Liu, PSU)	
13	Monday, 11/23/2020	Student presentation: Entity and Relation Extraction (Megan)	
13	Wednesday, 11/25/2020	Thanksgiving Holiday (No classes)	
14	Monday, 11/30/3030	Keyphase extraction (Krutarth Patel)	
14	Wednesday, 12/2/2020	Project time	
15	Monday, 12/7/2020	Final project presentation	
15	Wednesday, 12/9/2020	Final project presentation	Final report due by noon.

* Course schedules are subject to change depending on availability of speakers and the instructor.

Exam Schedule

No final exams.