University of North Carolina at Charlotte
DSBA 6188/ITIS 6010 Text Mining and Information Retrieval

Credits: 3 Credit Hours

Days, Time/Location:
9:00am -11:45am on Saturdays at Dubois Center (Uptown) 501

Course Description:
The availability of text data has created unprecedented opportunities to leverage computational and statistical approaches to turn data into actionable knowledge. This course covers general techniques for analyzing large amounts of text data as well as basic techniques for information retrieval.

The current technology of natural language processing has not yet reached a point to enable a computer to precisely understand natural language text, but text mining (TM) techniques with a wide range of statistical and heuristic approaches have been developed over the past few decades. They are usually very robust and can be applied to analyze and manage text data in any natural language, and about any topic. This course intends to provide a systematic introduction to many of these approaches, such as word association mining, topic modeling, and text classification. On the other hand, information retrieval (IR) is a relatively mature and well-established field. We will introduce the contemporary retrieval models as well as their evaluations.

We will offer Python code examples which contain implementations of many techniques discussed in this course. Homework exercises are designed based on Python to help students acquire practical skills of experimenting with the learned techniques and applying them to solve real-world application problems.

The required background knowledge to take this course is minimal since the it is intended to be mostly self-contained. However, students are expected to have basic knowledge about computer science, particularly some programming language, and be comfortable with some basic concepts in probability and statistics such as conditional probability and parameter estimation.

Faculty Information: Xi (Sunshine) Niu, Ph.D., Associate Professor
Office: Woodward 310G
Email: xniu2@uncc.edu
Office Hours: by appointment

Teaching Assistant: Mr. Chieh Wu
Email: cwu21@uncc.edu

Textbooks

Title: Text Data Management and Analytics: A Practical Introduction to Information Retrieval and Text Mining
Author(s): ChengXiang Zhai and Sean Massung
Publisher: ACM and Morgan & Claypool Publishers
Year: 2016

Title: Introduction to Information Retrieval
Author(s): Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze
Publisher: Cambridge
Year: 2008
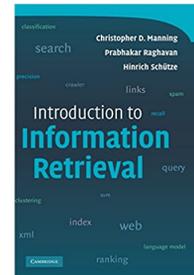
Title: Deep Learning for Matching in Search and Recommendation
Author(s): Jun Xu, Xiangnan He, Hang Li
Publisher: Now Publishers
Year: 2020
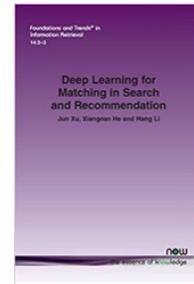
Evaluation Methods:

<u>Course grading will be based on these activities.</u>

Activities	Point
In-Class Quizzes	2 points x 14 = 28 points
After-Class Homework	5 points x 12 = 60 points
Term Project	12 points
Total	**101 points**

<u>Grade Scale:</u>
A = 90 points – 101 points
B = 80 points – 89 points
C = 70 points – 79 points
U = Below 70 points

Weekly Lesson Schedule:

Date	Contents

Jan 21	Syllabus Lesson 1: Basic Concepts
Jan 28	Lesson 2: Word Association Mining
Feb 4	Lesson 3: Topic Modeling I
Feb 11	Lesson 4: Topic Modeling II
Feb 18	Lesson 5: Text Classification
Feb 25	Lesson 6: Retrieval Model I: Boolean Retrieval
Mar 4	Spring Break, no classes
Mar 11	Lesson 7: Retrieval Model II: Vector Space Model 1
Mar 18	Lesson 8: Retrieval Model II: Vector Space Model 2
Mar 25	Lesson 9: Evaluations in Information Retrieval
Apr 1	Lesson 10: Retrieval Model III: Probabilistic Information Retrieval
Apr 8	Lesson 11: Retrieval Model IV: Web Search 1
Apr 15	Lesson 11: Retrieval Model IV: Web Search 2
Apr 22	Lesson 13: Overview of Deep Learning
Apr 29	Lesson 14: Retrieval Model V: Deep Learning Models for Information Retrieval

Course Policies:

Course Credit Workload:
This 3-credit course requires 9-12 hours effort (including the class time) for this course each week for approximately 14 weeks. Efforts may include but is not limited to: required reading, homework assignments, and studying for quizzes.

Class Attendance Policy:
Attending every class is mandatory. Class attendance entails being prepared, present, and attentive for the entire class period. Missing class reduces your grade through the following method: Two absences could be excused if you send an email with your explanation BEFORE the beginning of the class. More than two absences (three or above) in total will result in U in the course. For each absence, the student is responsible for catching up with all covered materials and assignments.

Late Submissions:
For assignments, unexcused late submission (according to the Canvas timestamp and the "late" flag) will receive a grade of 0. You should plan sufficiently for completing and submitting assignments. Should an emergency arise that greatly disrupts one's ability to complete an assignment, please send an email to Dr. Niu before the due date with a plan for submission after the due date. You need to receive Dr. Niu's permission for late submission.

Special Needs and Religious Accommodation:
If you have a documented disability and require accommodation in this course, contact the Office of Disability Services (https://ds.uncc.edu/students/academic) the first week of the semester. Accommodations for learning will be arranged by that office and communicated to the Instructor.

It is the obligation of students to provide faculty with reasonable notice of the dates of religious observances on which they will be absent by submitting a Request for Religious Accommodation Form to their instructor prior to the census date for enrollment for a given semester. The census date for each semester (typically the tenth day of instruction) can be found in UNC Charlotte's Academic Calendar (https://registrar.uncc.edu/printable-calendar).

Copyright and Permissions:
My lectures and course materials, including presentations, quizzes, homework problems and answers, and similar materials, are protected by copyright. I am the exclusive owner of copyright in those materials I create. I encourage you to take notes and make copies of course materials for your own educational use. However, you may not, nor may you knowingly allow others to reproduce or distribute lecture notes and course materials publicly without my express written consent. This includes providing materials to commercial course material suppliers such as CourseHero, Chegg, and other similar services. Students who publicly distribute or display or help others publicly distribute or display copies or modified copies of an instructor's course materials may be in violation of University Policy 406, The Code of Student Responsibility.

University Policies:

Code of Student Responsibility: https://legal.uncc.edu/policies/up-406
Code of Student Academic Integrity: https://legal.uncc.edu/policies/up-407
Diversity and Inclusion: https://diversity.uncc.edu/
Sexual Misconduct and Interpersonal Violence:
https://legal.uncc.edu/policies/up-502
Standard for Responsible Use: https://oneit.uncc.edu/iso/standard-responsible-use