University of North Carolina at Charlotte
College of Computing and Informatics
Department of Software and Information Systems

Course Number and Title: ITIS 6162: Knowledge Discovery in Databases

Credits, Days/Time, Location: 3 Grad Credits; Thursdays 3:30–6:15 p.m. in CHHS 159

Faculty Information: Xi Niu, Ph.D., Assistant Professor
Office: Woodward 310G
Office Hours: By appointment
Email: xniu2@uncc.edu (preferred way of contact)

Catalog Description – include Pre and/or Co-requisites:
The entire knowledge discovery process is covered in this course. Topics include: setting up a problem, data preprocessing, data mining in search for knowledge, knowledge evaluation, visualization and application in decision making.

Prerequisite(s): ITCS 6160 or permission of instructor.
Cross-listed as: DSBA 6162, HCIP 6162, ITCS 6162, ITIS 8162

Course Objectives:
About two thirds of the course content focuses on numeric data analytics and the other third text data analytics. Both R and Python will be used as the toolkits. Topics include: setting up a problem, data preprocessing, statistical analyses, pattern mining, classification, clustering, and several common text mining and analytics techniques. The main focuses of this course are both concepts and implementations/applications.

Teaching Strategies: Teaching methods include a combination of lectures for concepts, demos for R or Python implementations, and in-class quizzes for learning assessment, and activities for sharing ideas.

- 3:30pm ~ 4:20pm Lecture
- 4:20pm ~ 4:30pm Break
- 4:30pm ~ 5:20pm Lecture
- 5:30pm ~ 6:15pm Quiz & Reflections

Required Texts:
Title: Data Mining: Concepts and Techniques
Author(s):   Jiawei Han, Micheline Kamber, and Jian Pei
Edition: 3rd Edition
Publisher: Morgan Kaufmann
Year: 2011
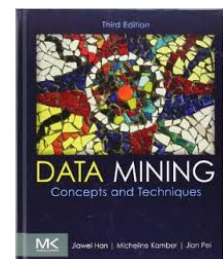
Title:	Text Data Management and Analytics: A Practical Introduction to Information Retrieval and Text Mining
Author(s):	Chenxiang Zhai and Sean Massung
Edition:	1st Edition
Publisher:	ACM and Morgan & Claypool Publishers
Year:	2016
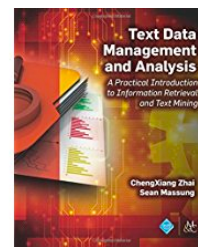
Title:	Discovering Statistics Using R
Author(s):	Andy Field, Jeremy Miles, and Zoe Field
Publisher:	Sage Publication Ltd
Year:	2012
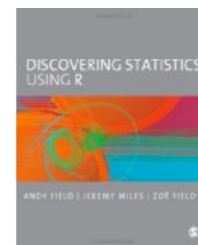
(The instructor will provide the electronic copies of all the textbooks)

Evaluation Methods:

Students should complete the weekly assigned readings **before** each class. At the end of each class, the students will be given a quiz. After the class, students need to finish the homework (if there is that week) and work on group assignments. During the semester, students will form groups of three and work on a mid-term toolkit demo and a final project.

<u>Mid-term toolkit demo</u>

For each group, please choose one of these following toolkits for the demo assignment

- Lucene
- Indri/Lemur
- META
- Weka
- Stanford NLP
- Scikit Learn
- NLTK
- Some Python library not covered in the course
- Some R package not covered in the course

For the selected toolkit, prepare a word version tutorial that includes:

- Brief introduction of the toolkit

- Overview of core functionality
- Demo of several functionality with some examples
- Description of the association with the course content

During the mid-term toolkit demo, each group has 10 minutes to present the word version tutorial (Powerpoint slides are not needed).

Final project

This is a comprehensive class project that needs the group to apply text analytics and data mining skills to a problem of "computational surprise". Detailed description of is attached in the Canvas assignment section.

During the final project presentation, each group has 10 minutes to present their project. A final report should be submitted one week after the presentation day incorporating the comments on the presentation day.

Course grading will be based on these activities.

Activities	Point
Class participation	10 points
Quizzes	3 points x12 = 36 points
Exercises	6 points x4 = 24 points
Midterm demo (group)	10 points
Final presentation (group)	10 points
Final report (group)	10 points

Grade Scale:
A = 90-100
B = 80 - 90
C = 70 - 80
U = Below 70

Topical Outline (Weekly Schedule):

Date	Lecture	Assignments
Lesson 1: Aug 24, Thursday	Syllabus Overview of Data Mining and Text Mining	
Lesson 2: Aug 31, Thursday	Getting to Know Your Data * Reading: Han Chapter 1 and Han Chapter 2	Quiz L2
Lesson 3: Sept 7, Thursday	Data Preprocessing Getting Familiar with R Getting Familiar with Python	Quiz L3

	* Reading: Han Chapter 3	
Lesson 4: Sept 14, Thursday	Correlation Linear Regression * Reading: Field Chapter 6.1-6.5 and Field Chapter 7.1-7.9	Quiz L4 Homework 1 due
Lesson 5: Sept 21, Thursday	Categorical Data Analysis (Chi-Square Test and Logistic Regression) * Reading: Field Chapter 8.1-8.8 and Field Chapter 18.1-18.6	Quiz L5
Lesson 6: Sept 28, Thursday	Word Association Mining * Reading: Zhai Chapter 12 and 13	Quiz L6 Homework 2 due
Lesson 7: Oct 5, Thursday	Topic Analysis I * Reading: Zhai Chapter 17.1-17.3	Quiz L7
Lesson 8: Oct 12, Thursday	Topic Analysis II * Reading: Zhai Chapter 17.4-17.6	Quiz L8
Midterm Day Oct 19, Thursday	Midterm Toolkit Demo Presentation	Midterm presentation slides due
Lesson 9: Oct 26, Thursday	Pattern Mining * Reading: Han Chapter 6	Quiz L9 Homework 3 due
Lesson 10: Nov 2, Thursday	Classification 1 * Reading: Han Chapter 8	Quiz L10
Lesson 11: Nov 9, Thursday	Classification 2 * Reading: Han Chapter 9	Quiz L11
Lesson 12: Nov 16, Thursday	Cluster Analysis * Reading: Han Chapter 10	Quiz L12
Nov 23, Thursday	Thanksgiving Break, no class	
Lesson 13: Nov 30, Thursday	Text Clustering & Categorization * Reading: Zhai Chapter 14 and 15	Quiz L13
Final Presentation Day Dec 7, Thursday	Final Project Presentation	Final presentation slides due
Final Day	No class	Homework 4 due

Dec14, Thursday		Final Report due

    ∗   All the due time is 3:30pm that day.

Course Policies

Attendance is mandatory. A basic requirement of this course is that you will participate in all class meetings, and complete all required course activities and assignments. Class attendance is required. It entails being present and attentive for the entire class period. Attendance shall be taken in every class.

Missing class reduces your grade through the following grade reduction policy: You are allowed two absences. Each additional absence, results in 3 points reduction from the class participation points. More than four (five or above) absences result in U in the course. For all absences, the student is responsible for all covered materials and assignments.

Absences must be explained with the email **sent to the instructor and cc'ed to the TA** before the beginning of a class.

Late submissions. For individual work, late submission (according to the Canvas timestamp and the "late" flag) **will receive a grade of 0**. Team members should plan sufficiently for completing group assignments. Should an emergency arise that greatly disrupts a team's ability to complete an assignment, permission and documentation must be received BEFORE the due date with a plan for submission after the due date.

Diversity Statement. No student will be discriminated against in the class based upon age, race, nationality, religion, sexual orientation, gender identity/expression, veteran's status, country of origin, or group affiliation. Likewise, all participants in this class will be expected to respect other members who fall into these categories. Any student who does not behave in a respectful manor with their classmates will be withdrawn from the class.

UNIVERSITY, COLLEGE AND DEPARTMENTAL POLICIES

University Policies:

Code of Student Responsibility:
"The *UNC Charlotte Code of Student Responsibility* (the Code) sets forth certain rights and responsibilities in matters of student discipline. The Code defines these responsibilities and guarantees you certain rights that ensure your protection from unjust imposition of disciplinary penalties. You should familiarize yourself with the provisions and procedures of the Code" (Introductory statement from the UNC Charlotte brochure about the Code of Student Responsibility). The entire document may be found at this Internet address: **http://legal.uncc.edu/policies/ps-104.html**

Academic Integrity:
All students are required to read and abide by the Code of Student Academic Integrity. Violations of the Code of Student Academic Integrity, including plagiarism, will result in disciplinary action as provided in the Code. Students are expected to submit their own work,

either as individuals or contributors to a group assignment. Definitions and examples of plagiarism and other violations are set forth in the Code. The Code is available from the Dean of Students Office or online at: http://www.legal.uncc.edu/policies/ps-105.html.

Integrity Statement is required for Homework 1 through Homework 4. At the beginning of each homework, there should be a statement: "I completed this Homework in compliance with the Code of Student Academic Integrity." Please sign your electronic signature after this statement.

<u>**Course Credit Workload:**</u>
This 3-credit course requires **three** hours of classroom with direct faculty instruction and **six** hours of out-of-class student work each week for approximately 15 weeks. Out-of-class work may include but is not limited to: required reading, resource research, written assignments, and studying for quizzes, and work on group projects.

<u>**Special Needs:**</u> If you have a documented disability and require accommodation in this course, contact Disability Services, Fretwell 230, phone: 687 4355 voice/TDD) the first week of the semester. Information about available services may be found at http://legal.uncc.edu/policies/ps-51.html. Accommodations for learning will be arranged by that office and communicated to the Instructor. If you speak English as a second language, please inform the instructor.

<u>**Diversity Statement:**</u>
UNC Charlotte strives to create an academic climate in which the dignity of all individuals is respected and maintained. Therefore, we celebrate diversity that includes, but is not limited to ability/disability, age, culture, ethnicity, gender, language, race, religion, sexual orientation, and socio-economic status.

All students are required to abide by the UNC Charlotte Sexual Harassment Policy (http://www.legal.uncc.edu/policies/ps-61.html) and the policy on Responsible Use of University Computing and Electronic Communication Resources (http://www.legal.uncc.edu/policies/ps-66.html). Sexual harassment, as defined in the UNC Charlotte Sexual Harassment Policy, is prohibited, even when carried out through computers or other electronic communications systems, including course-based chat rooms or message boards.

<u>**Religious Accommodation**</u>:
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 http://legal.uncc.edu/policies/ps-134.html . The census date for each semester (typically the tenth day of instruction) can be found in UNC Charlotte's Academic Calendar (http://registrar.uncc.edu/calendars/calendar.htm) .