INFO 4670 / INFO 5810 Data Analysis and Knowledge Discovery
Fall 2018

Contact Information

Instructor:	Dr. Jiangping Chen Professor Department of Information Science College of Information, University of North Texas 1155 Union Circle #311068 Denton, Texas 76203-5017
Office:	Discovery Park, Room E297J
Office hours:	By appointment -- arrange via course messages*
Email:	Jiangping.Chen@unt.edu
Voice:	Phone: (940) 369-8393

Teaching Assistant:	Pranathy Enamela PhD Student
TA's Email:	PranathyEnamela@my.unt.edu
Office:	Discovery Park, Room E297 J
Office Hour:	Mon, Thu 2pm to 4pm or by appointment -- arrange via course messages*

***Use UNT Canvas (https://unt.instructure.com/) Discussions and Course Messages tools for all course-related communication.**

Course Description

This course will introduce the student to data analysis, data mining, text mining, and knowledge discovery principles, concepts, theories, and practices. It is designed for the aspiring or practicing information professionals and covers the basics of working with data from a hands-on and practical perspective. Classes will incorporate lecture, discussion, the practice of learned concepts, and readings. The student will learn how to approach data and data mining tasks and techniques using Microsoft Excel and other data analysis tools through practice exercises and instructor-led training. They will learn the basic principles and theories of data mining, and text mining techniques as well as the business application of data mining and knowledge discovery tools.

Course Objectives

The learner will be an active and engaged participant in discussion forums within the learning community by analyzing, constructing/creating, and evaluating information presented within the textbook, external readings/resources, student research, and class activities.

The learner will be able to construct an appropriate bibliography in APA6 from scholarly sources of material for study and research.

The learner will be able to demonstrate an understanding of the fundamental principles, concepts, theories, and practices of data analysis, data mining, and knowledge discovery and discuss the interplay between them.

The learner will apply gained knowledge to solve real problems with datasets provided using skills developed in the course.

To achieve the above learning objectives, students are expected to study 9 - 12 hours per week for this course.

Prerequisites

Students are expected to have completed core courses, or be concurrently enrolled. It is the responsibility of the student to have a basic understanding of the Internet and general computing literacy.

Appointments and Online Meetings

Students are welcome to make an appointment with Dr. Chen and Pranathy to discuss course related questions. It is preferred that students send an email via a Canvas message to make an appointment. This course will have a website in Canvas (https://unt.instructure.com). ALL course materials will be uploaded to the class website in Canvas.

Required Textbooks

1. Winston, W. (2014). *Microsoft Excel 2013 Data Analysis and Business Modeling.* Microsoft Press, ISBN-10: 0735669139 ISBN-13: 978-0735669130
2. North, M. (2012). *Data Mining For The Masses.* Global Text Project, ISBN-10: 0615684378, ISBN-13: 978-0615684376
   - Available at: http://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf

Note new versions of the above textbooks are also available. It is encouraged that you purchase the version of the textbooks. They will be required as textbooks to replace current ones in Spring 2019.

Recommended for Excel beginners

1. Etheridge, D. (2010). *Excel Data Analysis: Your visual blueprint for creating and analyzing data, charts and Pivot tables.* (3rd ed.). Indianapolis: Wiley Publishing, Inc.
   - Available at: http://iii.library.unt.edu/record=b4558404~S12

Other class readings will be specified in related lessons.

Technical Requirements/Assistance

The following information has been provided to assist you in preparation for the technological aspects of the course.

You will need:
-Computer with **Microsoft Excel 2013 or above** and capable of accessing and using Canvas.

Undergraduate versus Graduate Course Requirements

The course requirements differ between what is required of undergraduate students versus what is required of graduate students and those differences are presented in the sections labeled Assignments (undergraduate) and Assignments (graduate). There are substantive differences in the quantity of work and the weighting of assignments. The student is responsible for understanding the requirements that apply to their own classification.

International Mater Students Holding F-1 Visa

International students who hold F-1 Visa must meet with the instructor or the TA in the F2F section. Please refer to the Catalog to identify the section number, date and time, as well as the meeting location.

Assignments and Assessments

Students are expected to study the 7 modules, participate in the discussion, and turn in assignments and project's deliverables on time. The final grade is an accumulation of the following:

Evaluation Criteria (undergraduate)
Discussion Board 20%
Learning Modules 50%
Midterm Exam 10%
Data Analysis Project 20%
TOTAL 100%

Discussion Board (20 %)
Students are expected to participate throughout the course in an active and interactive manner. Every week I will post two questions based on the readings assigned for the week. Use the assigned reading to respond in your post. Please follow the Rubric for Discussion Board Participation posted in the Rubric section of Course content.

Learning Modules (50%) – Due every two weeks, readings due as assigned
The course is comprised of seven learning modules. Each learning module has an overview, assigned readings, and supplementary readings, and there are hands-on exercises in 5 of the 7 modules. A description of each of the modules is included below.

Course Modules run for two-week periods and after the period elapses, the modules will be closed for submissions. Do not fail to turn in the hands-on assignments and forfeit points that represent a large portion of your grade!!!

Midterm Assessment (10 %) -due November 18th, by 11:59pm, through the Canvas assignment submission tool:
The midterm exam is a research paper. The paper will consist of a written response to a relevant question drawn from the course material and is to be the result of research conducted in the library. The student will select one question from a list of 5 possible questions provided by the instructor in the detailed instructions, shown below. The paper will be written using APA 6th Edition style guidelines, and a component of the scoring of your paper will be based on proper conformity to these standards.
A good reference for APA6 style guidelines is the Purdue OWL (online writing lab, found here https://owl.english.purdue.edu/owl/resource/560/01/ or consult a UNT librarian.

Detailed Midterm Instructions

- Undergraduates: Respond to the question you have chosen by writing a 5 to 7-page essay on the topic. The essay should be formatted as follows:

- Use APA 6th Edition for citations and references and any point not addressed in these instructions
- Minimum of 5 references.
- For all text, use 12-point Cambria (preferred) or 12-point Times New Roman font
- Text should be double spaced within paragraphs and double spaced between paragraphs
- Spell-check the document
- Do not include a title page or an abstract page
- Do not include References in the page count
- If you use section headings in the paper, use 12-point or larger type.
- 1-inch margins
- At top right on page 1, type three lines: your full name, course number, and semester with year.

Data Analysis Project (20%) – due December 10th, by 11:59pm, through the Canvas Assignment Submission link
The final project for the course will be a text mining assignment using the data mining software, Rapid Miner. The format and detailed instructions for the project will be provided prior to the distribution of the project.

Evaluation Criteria (graduate)
Discussion board 20%
Learning Modules 50%
Mid-Term Exam 10%
Data Analysis Project 20%
TOTAL 100%

Discussion Board (30 %)
Students are expected to participate throughout the course in an active and interactive manner. Every week I will post two questions based on the readings assigned for the week. As graduate students, you are expected to go beyond the course assigned readings, and support your posts with relevant additional references in APA style format in your responses. You can also use the assigned reading to make your posting. Please follow the Rubric for Discussion Board Participation posted in the Rubric section of the Course content.

Learning Modules (50%) – Due every two weeks, readings due as assigned
The course is comprised of seven learning modules. Each learning module has an overview, assigned readings, and supplementary readings, and there are hands-on exercises in 5 of the 7 modules. A description of each of the modules is included below.

Course Modules run for two-week periods and after the period elapses, the modules will be closed for submissions. Do not fail to turn in the hands-on assignments and forfeit points that represent a large portion of your grade!!!

Mid term Assessment (10 %) -due November 18th, by 11:59pm, through the Canvas assignment submission tool:
The midterm exam is a research paper. The paper will consist of a written response to a relevant question drawn from the course material and is to be the result of research conducted in the library. The student will select one question from a list of 5 possible questions provided by the instructor in the detailed instructions, shown below. The paper will be written using APA 6th Edition style

guidelines, and a component of the scoring of your paper will be based on proper conformity to these standards. A good reference for APA6 style guidelines is the Purdue OWL (online writing lab, found here
https://owl.english.purdue.edu/owl/resource/560/01/ or consult a UNT librarian.

Detailed Midterm Instructions
Graduates: Respond to the question you have chosen by writing a 10 to 15 page essay on the topic. The essay should be formatted as follows:

- Use APA 6th Edition for citations and references and any point not addressed in these instructions
- Minimum of 10 references.
- For all text, use 12-point Cambria (preferred) or 12-point Times New Roman font
- Text should be double spaced within paragraphs and double spaced between paragraphs
- Spell-check the document
- Do not include a title page or an abstract page
- Do not include References in the page count
- If you use section headings in the paper, use 12-point or larger type.
- 1-inch margins
- At top right on page 1, type three lines: your full name, course number, and semester with year.

Data Analysis Project (20%) – due December 10th, by 11:59pm, through the Canvas Assignment Submission link
The final project for the course will be a text mining assignment using the data mining software, Rapid Miner. The format and detailed instructions for the project will be provided prior to the distribution of the project.

The UNT scale for **grading** is as follows:

A = 90-100
B = 80-89
C = 70-79
D = 60-69
F = 59 and below

Late Submission

Students are expected to submit assignments and projects on time. **The due dates for discussion posts are Thursdays 11:59 pm and assignments are Saturdays 11:59 pm unless mentioned otherwise.** If an extenuating circumstance such as a medically diagnosed illness or family emergency arises, which prevents you from submitting your assignments, you should contact the instructor and the TA as soon as possible before the due date. **Late work without the permission of the instructor will receive a grade with a 10% penalty (or 10 points out of 100) per day after the due date. However, the late submission policy may apply on the prior circumstance.** A student who is having a trouble with the assignments is strongly encouraged to contact the instructor and the TA as early as possible for personal advising.

Incomplete

The UNT Graduate Catalog (http://catalog.unt.edu/index.php?catoid=19) describes and explains grading policies. A grade of Incomplete (I) will be given only for a justifiable reason and only if the student is passing the course. The student is responsible for meeting with the instructor to request an incomplete and discuss requirements for completing the course. If an incomplete is not removed within the time frame agreed to by instructor and student, the instructor may assign a grade of F.

Withdrawal

The UNT Graduate Catalog (http://catalog.unt.edu/index.php?catoid=19) describes and explains withdrawal policies and deadlines. The UNT semester course schedule lists specific deadlines regarding withdrawal. A grade of Withdraw (W) or Withdraw-Failing (WF) will be given depending on a student's attendance record and grade earned. Please note that a student who simply stops attending class and does not file a withdrawal form may receive an F

Academic Integrity

UNT has established a new policy on academic integrity, which can be found at the Provost office website: http://policy.unt.edu/sites/default/files/06.003.pdf. Specifically, UNT policy 18.1.16 (http://policy.unt.edu/sites/default/files/untpolicy/pdf/7-Student_Affairs-Academic_Integrity.pdf) defines the categories of academic dishonesty, including cheating, plagiarism, forgery, fabrication facilitating academic dishonesty, and sabotage. Also, it authorizes the instructor to determination that academic dishonesty has occurred. "A finding by an instructor that academic dishonesty occurred may be considered grounds for more serious academic penalties, up to and including failure in the course."

Penalties on Academic Dishonesty

- **First Time Violation.** The instructor will follow UNT procedure (http://policy.unt.edu/sites/default/files/06.003.pdf Section III B) and report the case to UNT Office for Academic Integrity (AIO). Student will receive "0" for the assignment/project that he/she performs academic dishonesty;
- **Second Time Violation.** The instructor will follow UNT procedure on multiple violation (http://policy.unt.edu/sites/default/files/06.003.pdf Section IV A) and report the case to UNT AIO. Student will receive "F" for this course.

Americans with Disabilities Act Compliance Statement

The Department of Library and Information Sciences, University of North Texas is committed to full academic access for all qualified students, including those with disabilities. In keeping with this commitment and in order to facilitate equality of educational access, faculty members in the Department will make reasonable accommodations for qualified students with a disability, such as appropriate adjustments to the classroom environment and the teaching, testing, or learning methodologies when doing so does not fundamentally alter the course.

If you have a disability, it is your responsibility to obtain verifying information from the Office of Disability Accommodation (ODA) and to inform me of your need for an accommodation. Requests

for accommodation must be given to me no later than the first week of classes for students registered with the ODA as of the beginning of the current semester. If you register with the ODA after the first week of classes, your accommodation requests will be considered after this deadline.

Grades assigned before an accommodation is provided will not be changed. Information about how to obtain academic accommodations can be found in UNT Policy 18.1.14, at www.unt.edu/oda, and by visiting the ODA in Room 321 of the University Union. You also may call the ODA at 940.565.4323.

Module	Topics	Readings	Release Time
1	**Data Analysis Introduction** Principles, Concepts, and Practices of Data Analysis **Hands On**: Excel introduction, Internet Resources, Library Resources, Entering data in a spreadsheet and creating charts and graphs.	Readings for Module 1 are located in Course Content, Module 1. They cover a variety of introductory concepts including formatting data for analysis, data types and display types, visual considerations etc.	**Week 1**
2	**Data Modeling Basics** Database and Data warehouse concepts **Hands-on**: Getting data into Excel and working with Excel	Book Chapters: 2 – Lookup Functions, 44 – The Data Model Reading Topics: Conceptual modeling of data, sets, tables, spreadsheets, and cubes	**Week 3**
3	**Operating on Data** Functions	Book Chapters: 3 – Index Function, 4 – Match Function, 5 – Text Functions, 6 – Date and Time Functions, 8,7 – Array Formulas and Functions Reading Topics: Functions and formulas	**Week 5**
4	**Knowledge Discovery Introduction** Topic: Principles, concepts, theories and practices of knowledge discovery Hands-on Topics: No topics.	Book Chapters: 13 – The Paste Special Command, 38 – Importing Data from a Text File or Document, 39 – Importing Data from the Internet, 40 – Validating Data Reading	**Week 7**

		Topics: Discovering knowledge in large databases	
5	Data mining fundamentals and basics of working with multi-dimensional data Topic: Semantic distance, Similarity, Classification, Clustering, Association, Regression, Online analytical processing - OLAP, ROLAP, MOLAP, and Cubes Hands-on Topics: Installing and exploring RapidMiner.	Book Chapters: 43 – Using PivotTables and Slicers to Describe Data Reading Topics: Data cleansing, text mining, similarity, dimension facts and measures	**Week 8**
6	**Visual Display of Data** Topic: Illustration, graphics, quantitative displays Hands-on Topics: Textual Analysis with RapidMiner.	Book Chapters: 23 – Conditional Formatting, 24 – Sorting in Excel, 25 – Tables, 41 – Summarizing Data by Using Histograms, 47 – Sparklines, 52 – Charting Tricks Reading Topics: Visual aesthetics in interface design	**Week 10**
7	**Advanced data and data mining techniques** Topic: Rule induction (if then else), Neural networks (trainable nets for decision making/prediction), Decision trees (classification), Genetic algorithms Hands-on Topics: No topic.	Book Chapters: 27- The Analytics Revolution, 28 – Introducing Optimization with Excel Solver, Optional – 29-36 more Solver Reading Topics: Types and examples of advanced data analysis	**Week 12**

Table 3. Study Schedule and Due Dates

Academic Week	Dates	Study Focus	Posting/Assignment/Project /Quiz Due Dates
1	Aug 27 – Sep 2	Syllabus Module 1	Self-Introduction (not graded) 2 Discussion posts due 30th Aug
2	Sep 3 – Sep 9	Module 1	Assignment 1 due 8th Sep
3	Sep 10 – Sep 16	Module 2	2 Discussion posts due 13th Sep
4	Sep 17 – Sep 23	Module 2	Assignment 2 due 22nd Sep
5	Sep 24 – Sep 30	Module 3	2 Discussion posts due 27th Sep
6	Oct 1 – Oct 7	Module 3	Assignment 3 due 6th Oct
7	Oct 8 – Oct 14	Module 4	2 Discussion posts due 11th Oct
8	Oct 15 – Oct 21	Module 5	2 Discussion posts due 18th Oct
9	Oct 22 – Oct 28	Module 5	Assignment 4 due 27th Oct
10	Oct 29 – Nov 4	Module 6	2 Discussion posts due 1st Nov
11	Nov 5 – Nov 11	Module 6	Assignment 4 due 27th Oct
12	Nov 12 – Nov 18	Module 7	Mid Term Assessment due 18th Nov
13	Nov 19 – Nov 25	Happy Thanksgiving	
14	Nov 26 – Dec 2	Module 7	2 Discussion posts due 29th Nov
15	Dec 3 – Dec 9	Work on Term Project	Work on Term Project
16	Dec 10 – Dec 14	Finishing the Term Project	Data Analysis Project due 10th Dec by 11:59pm