UNIVERSITY OF NORTH CAROLINA
CHARLOTTE
COLLEGE OF COMPUTING AND INFORMATICS
School of Data Science

DSBA 6188 | Text Mining & Information Retrieval | 3 Credits

Course Information

Course Number/Section	**DSBA 6188 U90**
Course Title/Semester	Text Mining & Information Retrieval / **Fall 2025**
Days & Times	*Wednesday at 5:30 pm – 08:15 pm*
Location	*Dubois, Room 601*

Contact Information

Instructor	Dr. Shih-Hsiung Chou
Email Address	schou6@charlotte.edu
Office Hours	Email or Zoom by appt.

COURSE PREREQUISITES

- Graduate/Ph.D. student standing or permission of instructor.
- Python programming skill
- Related but not required
  - Applied Machine Learning (DSBA 6165)
  - Artificial Intelligence and Deep Learning and AI (DSBA 6165)
  - Database Systems for Data Scientist (DSBA 6160)
  - Special Topics in DSBA - Applied LLM (DSBA 6010)

COURSE DESCRIPTION

This course offers a comprehensive journey into text mining and information retrieval, designed to equip students with the essential skills to analyze and extract insights from textual data. We begin with the fundamentals, covering the entire text mining pipeline from data acquisition and preprocessing to core Natural Language Processing (NLP) techniques. The curriculum emphasizes practical application, with a significant focus on building solutions for common tasks such as text classification, clustering, sentiment analysis, and topic modeling. Additionally, this course will incorporate Large Language Models (LLMs), whether locally hosted or cloud-based, for various text mining and information retrieval tasks. Through hands-on programming assignments, students will gain practical experience in designing and evaluating robust text analysis solutions, bridging the gap between theory and real-world application.

STUDENT LEARNING OUTCOMES

- Apply various text preprocessing and representation techniques.
- Understand the fundamentals of LLMs and their role in text analysis.

- Utilize LLMs for tasks such as sentiment analysis, summarization, topic modeling, and NER.
- Implement Retrieval-Augmented Generation (RAG) for enhanced information retrieval.
- Develop knowledge graphs and apply GraphRAG.
- Build practical text-mining solutions, including chatbots and automated data processing pipelines, using tools like Ollama, n8n, Snowflake Cortex, and Streamlit.
- Design and evaluate robust text analysis and information retrieval systems.

GRADING AND ASSESSMENT CRITERIA

30% Assignments
10% Attendance
20% Quizzes
40% Group Project

GRADING SCALE FOR COURSE

A 90-100 B 80-89 C 70-79 U 69 and below

Please note, that I will not round up to another grade level. For example, if you get a '89.9', it will be a B.

ATTENDANCE POLICY

Two absences could be excused if you send an email with your explanation BEFORE the beginning of the class.

LATE ASSIGNMENTS, TEST GRADES, AND GROUP PROJECT GRADES

Late Assignments (assignments submitted past the due date) will receive 5% off out of 100% for every day it is late without prior written approval with the instructor. Assignments over a week late can still receive a 50% so long as it is turned in prior to the final class date. Assignments never submitted or completed will receive a 0. Quizzes cannot be retaken without written approval from the instructor.

Group project grades are based on the group leader submitting assignments on time. Participation in group projects and assignments IS REQUIRED! Points can be taken off at the instructor's discretion due to lack of participation.

TEXTBOOK

Required: No required textbook.
Recommended:

- Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020), *Practical natural language processing: A comprehensive guide to building real-world NLP systems,* O'Reilly Media.
- Albrecht, J., Ramachandran, S., & Winkler, C. (2020). *Blueprints for Text Analysis Using Python: Machine Learning-Based Solutions for Common Real World (NLP) Applications*. O'Reilly Media.
- Boonstra, L. (2024). Prompt Engineering (Whitepaper). Google. https://www.kaggle.com/whitepaper-prompt-engineering

OTHER RESOURCES: Access to DataCamp and Snowflake will be provided free of charge to students enrolled in the class.
Students will create accounts using their **@charlotte.edu** email to access DataCamp. Instructor will provide accounts to students during the class.

HARDWARE:
The recommended computer specifications for a student are as follows:

- CPU: An Intel Core i5, AMD Ryzen 5, or Apple M4 processor, or a more powerful alternative.
- RAM: A minimum of 16GB.
- GPU: A dedicated GPU with at least 8GB of VRAM is required. Alternatively, an Apple M4 processor with its unified memory can be used.
- Storage: A 1TB Solid-State Drive (SSD).

SOFTWARE:
Students will be able to use MAC, Windows, or Linux. Students must have access to the internet. (required)

AI TOOL USE POLICY:
AI tools are only allowed to be used for projects and assignments; any usage for quizzes or exams is prohibited.

CLASS MEETING SCHEDULE:
The following class schedule and deadlines are subject to change at the discretion of the instructor and class circumstances. All assignments are due by the start of the next class which is generally Wednesday 5:30 pm unless otherwise indicated.

Date	Topic
August 20, 2025	**Week 1** ▪ Syllabus review ▪ Course Overview ▪ Introduction of text mining and information retrieval ▪ Install and configure required software including VS code, Ollama, n8n, Colab, and Snowflake account.
August 27, 2025	**Week 2** ▪ Text Preprocess—tokenization, stemming, stopwords, lemmatization, POS tagging, etc. ▪
September 3, 2025	**Week 3** ▪ Text Representations—BoW, TF-IDF, word embeddings, Vectorization, etc.
September 10, 2025	**Week 4**

	▪ Large Language Model Basic ▪ LLM tools—Ollama, Snowflake Cortex, n8n
September 17, 2025	**Week 5** ▪ LLM for ○ sentiment analysis ○ document summarization
September 24, 2025	**Week 6** ▪ LLM for ○ topic modeling ○ named entity recognition (NER)—langextract
October 1, 2025	**Week 7** ▪ LLM for ○ document search ○ Retrieval-Augmented Generation ▪ using n8n and Snowflake ▪ Final Project Preview
October 18, 2025	**Week 8** ▪ LLM for ○ knowledge graph ○ GraphRAG
October 15, 2025	**Week 9** Use Case—RAG Chatbot using Snowflake Cortex and Streamlit
October 22, 2025	**Week 10** Use Case—Local knowledge base development with RAG, n8n and Ollama
October 29, 2025	**Week 11** Use Case—Local knowledge base development with GraphRAG, LM Studio, and Ollama
November 5, 2025	**Week 12** Use Case—Automating the operations of an AI-powered news aggregation channel using n8n and Ollama
November 12, 2025	**Week 13** - Final Project Preparation or Question (on Zoom)
November 19, 2025	**Week 14** - Final project Presentation

November 26, 2025	**Week 15** ▪ Final Project Document Deliverable. Due 23:59. ▪ Happy Thanksgiving. No class.

PROJECT

Our course project will provide you the opportunity to explore and experience text mining collaborating with large language model in practice. You will collaborate with other students in this course as part of a group. The project will be assigned at the mid-point of the semester and each group will have the chance to choose between several projects provided by the instructor. A group can pitch an idea for an original project as well.

The project has several milestones in the form of project deliverables in order to keep your work progressing. Project deliverables must be met; no late work will be accepted. Students have the chance to correct deficiencies on their deliverables in all but the final project deliverable. **Participation is required. Peer reviews will be collected and made part of the project grade.**

READING

The readings for this course will be taken from the textbook and a variety of other current sources. Students must read the course materials and post any questions that you wish to be discussed on the forum.

GROUP DISCUSSION

The most vital use of Discussion Forums is to exchange ideas with other classmates. It is important that you check into the forums regularly. You are encouraged to ask questions regarding the required readings, discuss the unit topics, share information and resources with classmates, and respond to problems posted by your classmates or instructor. You should read everyone's posts and responses to the topics that interest you.

SUBMISSION OF WORK

Follow each assignment instruction; all work should be in PDF format or Jupyter notebook and uploaded into the assignment link as requested. Datacamp Assignments are graded in Datacamp. It is the students' responsibility to keep his/her copies of all work submitted to the instructor. All work are to be turned in by the due date, no late work will be accepted.

POLICY ON ACADEMIC INTEGRITY

The university policy 407, the Code of Student Academic Integrity, applies. This policy is available at http://legal.uncc.edu/policies/up-407.
Academic honesty is absolutely essential. Cheating, plagiarism or other academic misconduct will not be tolerated. If you are caught cheating, you will not pass this

course and will be subject to any and all penalties specified in the code of Student Academic Integrity. **If a student is found cheating, she or he will receive a ZERO for that assignment. If a student is found cheating a second time, she or he will receive an "F" for the course.**

Examples of violation academic integrity include, but are not limited to:

- pretending that somebody else's work is yours so that you can get a higher grade than your own work merits
- falsifying data
- lying in order to extend a deadline or gain some other special advantage
- helping other people to do any of these things
- copying answers on tests
- using prohibited reference materials (such as notes or books) during an exam
- turning in papers that you have not written yourself or that you wrote for a different course
- quoting material without marking it as quoted and without attributing it to its source (or closely paraphrasing material without attributing it to its source)
- misrepresenting a medical or family emergency or other personal contingency in order to delay a scheduled exam or to get extra time on an assignment
- pretending to have a disability you do not have (or exaggerating one you do have) in order to gain an unwarranted advantage unavailable to other students
- modifying graded material and then resubmitting it to "correct the error in grading"

RULES GOVERNING STUDENTS WITH SPECIAL REQUIREMENTS

Students with disabilities which require accommodations should:

1. Register with the Office of Disability Support Services and 504 Compliance to provide documentation
2. Bring the necessary information indicating the need for accommodation and what type of accommodation is needed. This should be done during the first week of classes or as soon as the student receives the information. If the instructor is not notified in a timely manner, retroactive accommodations may not be provided.

MISCELLANEOUS REQUIREMENTS

1. All requests to change grading of any course work must be submitted in writing within a week after the grades are made available. Requests must be specific and explain why you feel your work deserves additional credit.
2. All requests about missing (or zero) grades must be submitted in writing to the instructor within a week after the grades are announced. After that period the grade stands.
3. Please note that a student will not automatically receive an "I" grade when he/she misses some work, or misses the final exam. An "I" is given to those students who have a passing average at the time of the 'incident'. I grades must

go through a formal approval process and must be based on extenuating or emergency circumstances according to UNCC policy.

4. Submission of work: It is the student's responsibility to ensure that the instructor has received work submitted. This is especially important when work is submitted electronically.

   a. If you use email, ensure that you keep a copy of the sent email, and ask for a 'read receipt'.

   b. If submitting via our online course site Canvas, always keep a copy of your work.

5. Communication Protocol:

   (a) Questions, Comments, and Requests

   - For any questions or clarification of class material, please ask them on the Discussion Board in Canvas whenever possible. Everyone in the class is encouraged to help answer the questions. If satisfactory answers do not emerge, the instructor will answer.
   - For any comments or requests, please send email to the instructor

   (b) Canvas

   - Announcements will be posted in Canvas. Make sure to check the assignment area frequently enough to stay informed.
   - There are obviously things that are not appropriate for the Canvas discussion area, such as solutions for assignments (violation of honor code).

   (c) Emails

   - Each student is given an email account by UNC-Charlotte. This is the account that will be used by your instructor. Changes to class assignments or other course information will be posted online and may be sent to you. Check your email daily. Do not send email to your instructor from any other account, as it will be considered spam, and be deleted.
   - Please use Canvas, not emails, for general questions, unless you wish to keep your questions or comments private.
   - When emailing your instructor, please use a specific subject line starting with "DSBA- 6188: Homework 1 - [Last Name]."

STUDENT RESPONSIBILITIES

Please refer to University Policy 406 - The Code of Student Responsibility, http://legal.uncc.edu/policies/up-406, for specific information. In addition to the responsibilities specified by the University, for this course, it remains the student's responsibility to be aware of enrollment status, assignment due dates, changes to the syllabus, and deadlines for the UNCC academic calendar. Each student is responsible for his/her attendance and properly withdrawing from the course if necessary.

DISCLAIMER

This syllabus is intended to give the student guidance in what may be covered during the semester and will be followed as closely as possible. However, the instructor reserves the right to modify, supplement and make changes as needed.

Good luck in class! I am looking forward to working with you this Fall!