

University of North Carolina at Charlotte

DSBA6188/ITIS6010 Text Mining and Information Retrieval

Semester: Spring 2024

Credits: 3 Credit Hours

Days, Time/Location: 5:30 - 8:15 pm on Tuesdays at The Dubois Center (Uptown) 1101

Instructors: Dr. Chang Hsin Lee <clee184@charlotte.edu>

Dr. Ryan Wesslen <rwesslen@charlotte.edu>

Teaching Assistant: Neha Lahri <nlahri@charlotte.edu>

Course description:

In this course, we will explore the evolving landscape of natural language processing (NLP), examining how recent advancements and industry progress in applied NLP have transformed traditional concepts of information retrieval and text mining. While classical topics like information extraction and word embeddings will be covered, we will also occasionally delve into the pivotal role of large language models (LLMs), such as ChatGPT, in enhancing NLP workflows.

Guided exploration: Think of this course as a journey where you'll uncover "gold nuggets" of knowledge that go beyond what can be learned from videos or online courses. As your guides, we'll impart industry best practices, bring in guest lecturers with real-world experience, and help you discover practical applications of NLP in your current or future job.

Collaborative learning: Collaborate with fellow students (e.g., Slack and in-class), develop communication skills, work on hands-on programming tasks, and leverage LLMs to enhance productivity. Homeworks (done in groups of 2) and final project (groups of 3-4) will be in Canvas-assigned groups.

Critical evaluation: Learn to appropriately evaluate and critique NLP models, emphasizing practical skills that are crucial in real-world applications like reproducibility, version control (e.g., git), and models for production rather than only for development. Homework will resemble take-home [technical interviews](#).

Project-based: Immerse yourself in the world of code. We'll teach you how to read and evaluate code using tools like Jupyter/Colab notebooks in class. There will be limited lecture slides, and instead, we'll focus on hands-on activities and collaborative projects.

No Required (but recommended) textbooks: In line with the dynamic nature of the field, there are no required textbooks. We do provide optional textbooks that provide additional insights, but we'll only cover parts of them so it doesn't make sense to require them. However, we highly recommend them:

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. ([Jupyter examples](#))

Tunstall, L., Von Werra, L., & Wolf, T. (2022). [*Natural language processing with transformers*](#). "O'Reilly Media, Inc.". ([Jupyter examples](#))

Pre-requisite skills:

- Python core skills (lists, dictionaries, virtual environments)
- Command line (terminal, PowerShell, bash, zsh, or some other tool)
- Proficiency in a Python IDE like [Colab](#) or [VSCode](#)

> The class examples will be developed using a Mac/Linux (UNIX). While possible to use Windows OS, some course code may need to be modified for Windows. We **recommend** students with Windows [install Ubuntu through WSL2](#).

Grades:

In-class attendance & participation (10 pts, 10%)

- In-class attendance will be taken on a sign-in sheet every class during the first break.
- Permitted two (2) absences. Each additional absence will reduce 0.5 point.

> A Zoom link is offered in each class for students who have **excused** absences (e.g., positive COVID test). However, attending via Zoom will count as an absence. Zoom calls will **not** be recorded.

- Participation is based on in-class participation and/or Slack (e.g., helping other students)

> The course will use Slack as the primary communication tool. Click [here](#) to join. We'll use the free version which means messages are **only** kept for the **last 90 days**.

Calmcode applications (5 x 2 pts = 10 pts, 10%)

- Watch [calmcode](#) videos to expand your programming skills, and submit an artifact (e.g., notebook, repo, or script) that applies code from the video.
- Possible +0.5 bonus points for *excellent* submissions.

- Due dates: Every 3 weeks (but students can submit early).
- Late submissions will be penalized 50%.

Homeworks “take home interviews” (4 x 5 pts = 20 pts, 20%)

- Late submissions will be penalized 50%
- Can work with 1 other student; only submit once

Course project (40 pts, 40%)

- Group projects with 2 or 3 other students
- 10 points: 4 milestone check-ins
- 15 points: [PechaKucha](#) in-class presentation on **April 23**
- 15 points: deliverable on **May 7** (final)

Final exam “technical interview” (20 pts, 20%)

- in-class on May 7
- Cumulative, closed book (no notes), multiple choice and short answer

Administrative:

Orderly and productive classroom conduct

We will conduct this class in an atmosphere of mutual respect. We encourage your active participation in class discussions. Each of us may have strongly differing opinions on the various topics of class discussions. The conflict of ideas is encouraged and welcomed. The orderly questioning of the ideas of others, including mine, is similarly welcome. However, we will exercise my responsibility to manage the discussions so that ideas and arguments can proceed in an orderly fashion. You should expect that if your conduct during class discussions seriously disrupts the atmosphere of mutual respect we expect in this class, you will not be permitted to participate further.

Recording in the classroom

Electronic video and/or audio recording is not permitted during class unless the student obtains permission from the instructor. If permission is granted, any distribution of the recording is prohibited. Students with specific electronic recording accommodations authorized by the Office of Disability Services do not require instructor permission; however, the instructor must be notified of any such accommodation before recording. Any distribution of such recordings is prohibited.

Discussion of grades and performance

Such discussion shall occur between the student and the instructor(s). Sharing information regarding grades and performance in places such as discussion forums or email blasts is prohibited.

Code of Student Responsibility

“The purpose of the Code of Student Responsibility (the Code) is to protect the campus community and to maintain an environment conducive to learning. University rules for student conduct are discussed in detail. The procedures followed for any Student, Student Organization or Group charged with a violation of the Code, including the right to a hearing before a Hearing Panel or Administrative Hearing Officer, are fully described.” (Introductory statement from the UNC Charlotte brochure about the Code of Student Responsibility). The entire document may be found at this site: <https://legal.uncc.edu/policies/up-406>

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 work, either as individuals or as 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: <https://legal.uncc.edu/policies/up-407>.

Using Large Language Models (LLMs) for Homework and Assignments


Using Large Language Models like ChatGPT or CoPilot in your coursework (homework or projects) is permissible, provided you attribute by stating the tool, prompt, and any modifications you made. Cite AI assistance in assignments and summarize the original prompts and instructions used. Grades will be based on answer accuracy and the quality of your AI usage.



Save Course Dates via Feed to Google Calendar

Calendar Key:

 = Guest Lecturer

 = Class will be virtual

 = Guest Lecture will be outside of normal class hours

 The course calendar and times are subject to change. We'll use Canvas and Slack as the official forum for announcements and changes 

Date	Details	Due
Tue Jan 16, 2024	Calendar Event Class 1: Introduction	5:30pm to 8:15pm

Tue Jan 23, 2024	Calendar Event Class 2: Named Entity Recognition (NER)	5:30pm to 8:15pm
	Calendar Event Class 3: Word Embeddings	5:30pm to 8:15pm
Tue Jan 30, 2024	Assignment Calmcode 1	due by 11:59pm
	Assignment Project Milestone 1	due by 11:59pm
Tue Feb 6, 2024	Calendar Event Class 4: Using LLM's for Information Extraction 🧑	5:30pm to 8:15pm
Wed Feb 7, 2024	Assignment Homework 1: Reddit Cooking NER and Data Cleanup	due by 11:59pm
Tue Feb 13, 2024	Calendar Event Class 5: Transformers 🧑	5:30pm to 8:15pm

Calendar Event
Class 6: Applied NLP 5:30pm to 7pm
Tricks 🧑🌐🕒
Tue Feb 20, 2024

Assignment
Calmcode 2 due by 11:59pm

Calendar Event
Class 7: Transformer Applications 5:30pm to 8:15pm
Tue Feb 27, 2024

Assignment
Homework 2 due by 11:59pm

Assignment
Project Milestone 2 due by 11:59pm

Calendar Event
No Class: Spring Break 12am
🌴
Tue Mar 5, 2024

Calendar Event
Class 8: Search & Document Similarity 🧑
5:30pm to 8:15pm
Tue Mar 12, 2024

Assignment
due by 11:59pm

	Calmcode 3	
	Calendar Event	
Tue Mar 19, 2024	Class 9: Prompt Engineering 	5:30pm to 8:15pm
	Calendar Event	
	Class 10: Vector Databases & RAG   	5:30pm to 7pm
Tue Mar 26, 2024	Assignment	
	Homework 3	due by 11:59pm
	Assignment	
	Project Milestone 3	due by 11:59pm
	Calendar Event	
Tue Apr 2, 2024	Class 11: Low Resource NLP	5:30pm to 8:15pm
	Assignment	
	Calmcode 4	due by 11:59pm
Tue Apr 9, 2024	Calendar Event	5:30pm to 8:15pm

Class 12: Ethics &
Evaluation in NLP 🧑

Calendar Event

Class 13: Project
Check-in

5:30pm to 8:15pm

Tue Apr 16, 2024

Assignment

Calmcode 5

due by 11:59pm

Assignment

Project Milestone 4

due by 11:59pm

Assignment

20x20 (PechaKucha)
Presentation

due by 5:29pm

Tue Apr 23, 2024

Calendar Event

Class 14: 20x20 Project
presentations

5:30pm to 8:15pm

Assignment

Homework 4

due by 11:59pm

Tue Apr 30, 2024

Calendar Event

No class: Reading Day

12am

Tue May 7, 2024

Assignment

Final Exam: "Technical
Interview"

due by 5:29pm

Assignment

Final Project Deliverable

due by 5:29pm
