CLS 162/DH 120: Natural Language Processing and the Human Record

Spring 2025 (Full Semester)

Class meetings: Monday and Wednesdays 3:00 - 4:15 PM, PLACE TBD

Instructor: Peter Nadel, peter.nadel@tufts.edu

Instructor Office Housrs: *online*, visit this link to set up a meeting with me

Prerequisite: CLS 161/DH 101: Introduction to Digital Humanities is a prerequisite for this course. Students with an intermediate background in Python and experience in the Humanities may be admitted with instructor permission. Students are also encouraged to take CLS 160/DH 102: Quantitative Text Analysis.

Semester Hour Units (SHUs): 3 credits

Course overview

Welcome to Natural Language Processing and the Human Record! This course is the seminar-level offering from the Digital Humanities (DH) minor. A predominant theme in the course will be the study and analysis of *under-resourced* languages: languages, historical or contemporary, which have limited technological support, like grammars, parsers and machine-translation. As natural language processing (NLP) becomes more and more critical scholarship, these languages tend to be crowded out by well-resourced counterparts like English, French and Mandarin Chinese. In this course, we will look at new ways of addressing these problems.

To begin, we will refresh your understanding of dependency parsing and Python skills by exploring the advanced features of the software package `spaCy`. We will then turn to webscraping techniques, including data collection, curation and storage, to build a corpus of several non-English languages. Starting in week 8, we are going to start training out own neural nets and deploying these models to help us understand and interpret texts in under-resoursed languages. By week 11, we will explore advanced model training with machine translation, transformer-based architectures. This material will help us as we study and utilize cutting-edge tools like large language models (LLMs), artificial intelligence (AI) and retrieval augmented generated (RAG).

Students will be responsible for a final project which is meant to apply any of the technqiues that we studied to an under-resourced language of their choice.

Textbooks and readings: There is no textbook for this course. Instead, most class material will be posted on our Canvas page when appropriate. As NLP and DH are very interdisciplinary fields, we will also be reading several articles, from a variety of fields, which will help us think through different perspectives and technical reasonings.

Use of Artificial Intelligence (AI) tools such as ChatGPT: If you use AI tools to assist you in any part of the assignments or final project of the course, you must document it. Either in a separate note or included in the assignment write-up, tell me what tool you used and how you used it. You are not expected to use these tools for the course and students can succeed without use of these tools. I would encourage students to forgo use of these tools until week 11 and the case study section of the course at which point explaination of confusing code or model architecture could be helpful. Use of these tools is strongly discouraged for reading assignments.

Academic Conduct: Tufts holds its students strictly accountable for adherence to academic integrity. The consequences for violations can be severe. It is critical that you understand the requirements of ethical behavior and academic work as described in Tufts' Academic Integrity handbook. If you ever have a question about the expectations concerning a particular assignment or project in this course, be sure to ask me for clarification. Each student is responsible for upholding the highest standards of academic integrity, as specified in the Tufts University policies. It is the responsibility of each student to understand and comply with these standards, as violations will be sanctioned by penalties ranging from failure on an assignment and the course to dismissal from the school.

Grading and Course expectations: * Assignments: 50% * Final Project: 30% * Class participation: 20% *Nota Bene: Class participation consists of class attendance, office hours and engagement with course material. It is meant to give credit for your interest and enthuiasm with the course material. It is expected that all students will view the required videos, read the required articles, cases, chapters and other materials, and participate in online discussion forums. I reserve the right to adjust a grade if these requirements are not fulfilled.*

Grading range: A passing grade in the course is C- or better. Course grades will be based on the below ranges (subject to revision during the course):

Grade	Range
A	> 94%
A-	90 - <94%
B+	87 - <90%
B	84 - <87%
B-	80 - <84%
C+	77 - <80%
C	73 - <77%
C-	70 - <73%

Submission of Graded Material: Assignment due dates will be specified on Canvas. Assignments received after their deadline will be penalized five points per two days unless extension is approved in advance. Students who are unable to complete an assignment or exam on time for any reason should notify the instructor by email prior to the deadline, with a brief explanation for why the extension is needed.

Accommodation of Disabilities: Tufts University is committed to providing equal access and support to all students through the provision of reasonable accommodations so that each student may access their curricula and achieve their personal and academic potential. If you have a disability that requires reasonable accommodations please visit: https://students.tufts.edu/student-accessibility-services/how-we-help/academic-accommodations to make arrangements for determination of appropriate accommodations. Please be aware that accommodations cannot be enacted retroactively, making timeliness a critical aspect for their provision.

Diversity Statement: We believe that the diversity of student experiences and perspectives is essential to the deepening of knowledge in this course. We consider it part of our responsibility as instructors to address the learning needs of all of the students in this course. We will present materials that are respectful of diversity: race, color, ethnicity, gender, age, disability, religious beliefs, political preference, sexual orientation, gender identity, socioeconomic status, citizenship, language, or national origin among other personal characteristics.

Course Topics and Assignments by Week:

Week	Course Topics	Assignments Due
1	• **Syllabus overview** • Understanding NLP • Text classification	• Getting set up with Google Colab
2	• **Token classification** • Using pretrained model for Named Entity Recognition (NER) • Software used: `spaCy`	• **Assignment 1 assigned** - *Due Week 4*
3	• **Token classification** • Using pretrained model for treebanking • Software used: `spaCy`	• **Read**: Beyond Translation: Language Hacking and Philology
4	• **Webscraping** • Static webscraping • Software used: `BeautifulSoup` and `pandas`	• **Assignment 1 due** • **Assignment 2 assigned** - *Due Week 6*
5	• **Webscraping** • Dynamic website design and data collection • Software used: `selenium` and `pandas`	• **Read**: Fading Away... The challenge of sustainability in digital studies

Week	Course Topics	Assignments Due
6	• **Token classification** • Manual annotation for token classification • Software used: `Doccano`	• **Assignment 2 due** • **Assignment 3 assigned** - *Due Week 8*
7	• **Token classification** • Training new models for token classification • Software used: `spaCy`	• **Read** Developing Geographically Oriented NLP Approaches to Sixteenth–Century Historical Documents: Digging into Early Colonial Mexico
8	• **Feature Extraction** • Using word2vec for NER • Software used: `gensim` and `scikit-learn`	• **Assignment 3 due** • **Assignment 4 assigned** - *Due Week 10*
9	• **Feature Extraction** • Training word2vec from scratch • Software used: `numpy`	• **Read:** Efficient Estimation of Word Representations in Vector Space
10	• **Recurrent Neural Nets** • Case study: Machine Translation • Software used: `Pytorch`	• **Assignment 4 due** • **Assignment 5 assigned** - *Due Week 12*
11	• **Transformers** • Case study: Decoder-only Architecture, Generative Pretrained Transformers (GPTs) • Software used: `Pytorch`	• **Read** Language Models are Unsupervised Multitask Learners Pages 1-10
12	• **Transformers** • Case study: Encoder-only Architecture, Bidirectional Encoder Representations from Transformers (BERT) • Software used: `Pytorch`	• **Assignment 5 due** • **Final Project assigned** - *Due Week 14*
13	• **Tokenizers** • Case study: Byte-Pair Encoding (BPE) Tokenizers • Software used: `HuggingFace`	• **Read** wHy DoNt YoU jUsT uSe ThE lLaMa ToKeNiZeR??
14	• **Large Language Models** • Case study: Retrieval Augmented Generation • Software used: `HuggingFace`	• **Nothing due**; work on your final projects
15	• **Presentations**	• **Finish Final Project** - Due TBD