Natural Language Processing: Syllabus

Alan W. Black & David R. Mortensen
Carnegie Mellon University

Spring 2021

Instructors:	Prof. Alan W Black (awb@cs.cmu.edu) and David R. Mortensen (dmortens@cs.cmu.edu)
Teaching assistants:	TBA
Lecture time:	Tuesdays & Thursdays, 4:00–5:20
Location:	TBA
Web page:	http://demo.ark.cs.cmu.edu/NLP/
Faculty office hours:	By appointment (Black); By appointment (Mortensen)
TA Office hours:	TBA

1 Summary

This course is about a variety of ways to represent human languages (like English and Chinese) as computational systems, and how to exploit those representations to write programs that do useful things with text and speech data, like translation, summarization, extracting information, question answering, natural interfaces to databases, and conversational agents.

This field is called Natural Language Processing or Computational Linguistics, and it is extremely multidisciplinary. This course will therefore include some ideas central to Machine Learning (discrete classification, probability models) and to Linguistics (morphology, syntax, semantics).

We'll cover computational treatments of words, sounds, sentences, meanings, and conversations. We'll see how probabilities and real-world text data can help. We'll see how different levels interact in state-of-the-art approaches to applications like translation and information extraction.

From a software engineering perspective, there will be an emphasis on rapid prototyping, a useful skill in many other areas of Computer Science. In particular, we will introduce some high-level formalisms (e.g., regular expressions) and tools (e.g., Python) that can greatly simplify prototype implementation.

2 Target

The course is designed for SCS undergraduate students, and also to students in graduate programs who have a peripheral interest in natural language, or linguistics students who know how to program. Prerequisite: Strong programming abilities in Python and a knowledge of data structures and algorithms.

3 Evaluation

Students will be evaluated in five ways:

Exams (30%) two midterm exams, (each 15%)

Project (30%) a semester-long 4-person team project (see below).

Homework assignments (25%) 6 assignments, mostly small programming assignments, given roughly every week for the first portion of the semester.

Quizzes (15%) 10 Canvas quizzes given at the end of most weeks.

The lowest 2 homework grades and the lowest 3 quiz grades will be dropped. This policy is provided in lieu of a late-days or grace-days policy.

Late Policy No work will be accepted late.

Academic Honesty Exams and quizzes are to be completed individually. Verbal collaboration on homework assignments is acceptable, but (a) you must not share any code or other written material, (b) everything you turn in must be your own work, and (c) you must note the names of *anyone* you collaborated with on each problem (the *only* exceptions are the instructors and TAs), and the nature of the collaboration (e.g., "X helped me," "I helped X," "X and I worked it out together."). If you find material in published literature (e.g., on the Web) that is helpful in solving a problem, you must cite it and explain the answer in your own words. The project is to be completed by a team; you are not permitted to discuss any aspect of your project with anyone other than your team members, the instructor, and the TAs. You are encouraged to use existing NLP components in your project; you must acknowledge these appropriately in the documentation.

Suspected violations of these rules will be handled in accordance with the CMU guidelines on collaboration and cheating (`http://www.cmu.edu/policies/documents/Cheating.html`).

4 Project

A major component will be a 4-person team project. The project involves two parts:

- a **questioning** program (`ask`) whose input is a web page P and whose output is a set of questions about the content in P that a human could answer if she read P, and
- a **answering** program (`answer`) whose input is a web page P and a question Q about P and whose output is an intelligent answer A.

Projects will be pitted against each other in a competition at the end of the course. Here's how the competition works:

1. Questions will be generated automatically by the questioning systems developed by the teams.
2. This questions will be evaluated for correctness and fluency by member of the class in a blind setup.

3. A subset of the best questions will be selected programmatically by the TAs and used as test cases for the answering systems.

4. These questions will be passed to the answering systems.

5. The answers will be manually evaluated by members of the class in a blind setup for correctness, fluency, and conciseness.

The project will be primarily graded based on documentation your team submits describing how the programs work, and a brief video presentation at the end of the semester.

5 Textbook

The textbook for the course will be the second edition of *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, by Daniel Jurafsky and James H. Martin. The course will cover roughly sections I, III, IV, and parts of V.

6 Schedule

02/02	Course overview; What does it mean to know language?	1	
02/04	Information extraction, question answering, and NLP in IR	22.0–2, 23.0–2	
02/09	Project		HW1 due
02/11	Words, morphology, and lexicons	3.1–3.9	
02/16	Language models and smoothing	4.3–8	Initial plan due
02/18	Noisy channel models and edit distance	3.10–11, 5.9	
02/25	Part of speech tags	5.0–3	HW2 due
03/02	Sequence tagging		
03/04	Classification 1		HW3 due
03/09	Classification 2		
03/11	Deep Learning		HW4 due
03/16	Syntactic representations of natural language	12.0–3	
03/18	Chomsky hierarchy and natural language	15	Progress report due
03/23	Midterm I		
03/25	Treebanks	12.4, 14.7	HW5 due
03/30	Lexical semantics	17.0–2, 19.0–3	
04/01	Word embeddings/vector semantics	6 (SLP3)	
04/06	Contextualized representations (BERT)		
04/08	Verb/sentence semantics	17.2–4, 19.4–6	HW6 due
04/13	Compositional semantics, semantic parsing	18.1–3	
04/20	Discourse, entity linking, pragmatics	20.0–6, 20.8–11	
04/22	Speech 1		
04/27	Speech 2		Project dry run code due
04/29	Sequence-to-sequence models		
05/04	Machine Translation	25.0–1, 25.9	Final project code due
05/06	Midterm II		Final project report due

For a more complete schedule, including assignments, deadlines, slides, and videos, please see `http://demo.ark.cs.cmu.edu/NLP/`.

7 Remote Instruction

This class will be taught remotely in its entirety. It is suggested that all students attend lectures synchronously via Zoom, so they can answer questions and interact with the teaching staff. However, all lectures will be recorded and posted on Youtube, for later reference or in case you were not able to join synchronously on a particular day.

8 Technology Requirements

The following technologies are used in this course:

- Zoom (`https://zoom.us/`), for conducting live recitations
- Youtube (`https://www.youtube.com`, for distributing recorded lectures
- Piazza (`https://www.piazza.com`), for disseminating course information and answering student questions
- Gradescope (`https://www.gradescope.com`), for grading student homework and administering exams
- Canvas (`https://canvas.cmu.edu`), for providing access to the gradebook, handing in group work, and collecting private links

We recognize that the large number of technologies used in this course may be confusing to some students (for example, they might not know what gets handed in via Canvas versus Gradescope). Problematically, though, none of these tools offers all of the functionality the we need for the class as a whole. This combination has worked well in the past.

9 Students with Disabilities

If you have a disability and have an accommodations letter from the Disability Resources office, I encourage you to discuss your accommodations and needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at `access@andrew.cmu.edu`.

10 Student Wellness

We are all going through a very difficult year. The novel coronavirus has disrupted much of our lives and social distancing has left many of us feeling isolated and alone.

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

All of us benefit from support during times of struggle. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is almost always helpful.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at `http://www.cmu.edu/counseling/`. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

11 Statement on Diversity

We must treat every individual with respect. We are diverse in many ways, and this diversity is fundamental to building and maintaining an equitable and inclusive campus community. Diversity

can refer to multiple ways that we identify ourselves, including but not limited to race, color, caste, national origin, language, sex, disability, neurotypicality, age, sexual orientation, gender identity, religion, creed, ancestry, belief, veteran status, or genetic information. Each of these diverse identities shape the perspectives our students, faculty, and staff bring to our campus. We at CMU will work to promote diversity, equity and inclusion not only because diversity fuels excellence and innovation, but because we want to pursue justice. We acknowledge our imperfections while we also fully commit to the work, inside and outside of our classrooms, of building and sustaining a campus community that increasingly embraces these core values.

Each of us is responsible for creating a safer, more inclusive environment. Unfortunately incidents of bias or discrimination do occur, whether intentional or unintentional. They contribute to creating an unwelcoming environment for individuals and groups at the university. Therefore, the university encourages anyone who experiences or observes unfair or hostile treatment on the basis of identity to speak out for justice and support, within the moment of the incident or after the incident has passed. Anyone can share these experiences using the following resources:

- Center for Student Diversity and Inclusion: `csdi@andrew.cmu.edu`,
- (412) 268-2150 Report-It online anonymous reporting platform: `www.reportit.net` username: `tartans` password: `plaid`

All reports will be documented and deliberated to determine if there should be any following actions. Regardless of incident type, the university will use all shared experiences to transform our campus climate to be more equitable and just.