syllabus

schedule

WEEK 1

Introduction and Applications

January 12

- **Language** - the most efficient and compact way to transfer knowledge is through words, where the window to AGI is through NLP. This lecture is an introduction that takes us through history of how we got to LLMs. We'll also review some applications of NLP, current industry standards, and some of the most impactful approaches and where they are being implemented. Finally, we'll preview what we'll be learning, the logistics of how we'll be doing so, and the expectations for your participation in this class.
- Applications Overview
  - Machine Translation (Baidu's Word-Word)
  - Summarization (Dialogues, Newspaper Articles, etc.)
  - Text Classification and Clustering (News Article Groupings, etc.)
  - Question and Answering (LLMs and Chatbots)
- Submissions
  - Laboratory - Getting Started on Google Cloud with Your Credits
  - Assignment 1 is assigned - *A First Look at Processing Language*

WEEK 2

ML Foundations and Software Engineering

January 19

- As NLP is a specific branch of machine learning, we will review some foundational knowledge that we'll utilize through the course of this class. We'll look at both machine learning and software engineering best practices that will help you build and scale NLP systems later in the course. Because most NLP algorithms today rely heavily on computing resources, we'll dive into distributed compution approaches and cloud-based operations.
- Lecturing Topics
  - Foundations of Machine Learning
  - Software Engineering Practices
- Required Keynote Reading
- Submissions
  - Laboratory - Containerization in the Cloud
  - Assignment 1 is **due**
  - Assignment 2 is assigned - *Text Classification*

WEEK 3

Language Classification

January 26

- Building upon our review of machine learning, we discuss strategies in feature extraction and generation. Particularly as creating a vocabulary can explode required memory space, our featurization includes NLP-specific techniques (e.g., tokenization, lemmatization, etc.). This week also marks the first week of **required keynote paper** reading, where we will begin the tour of seminal papers that have revolutionized not only language processing but also machine learning and artificial intelligence writ large.
- Lecturing Topics
  - Building Vocabulary with Stopwords and Stemming
  - Preprocessing - Tokenization and Lemmatization
  - Logistic Regression Classifier
  - Naïve Bayes Classifiers
- **Application** - Sentiment Analysis
- Required Keynote Paper - GPT-5 Model Card from OpenAI
  - Older model reference - OpenAI's GPT-4 technical report

- Laboratory - Naïve Bayes

WEEK 4

Text Processing Algorithms

February 2

- One of the most widely used algorithms in practice today are autocorrecting algorithms that typically have on-device requirements. In this lecture, we'll review elements of dynamic programming, particularly with respect to the minimum edit distance algorithm, and how we can apply these concepts to the autocorrect and subsequently the autocomplete problem.
- Lecturing Topics
  - Representations of Language
  - Comparisons / Differences in Language
  - Minimum Edit Distance Algorithms
- Application - Autocorrect in Practice
- Required Keynote Paper - Outrageously Large Neural Networks
  - DeepSeek Technical Report - Mixture of Experts
  - Hugging Face Blog
- Submissions
  - Laboratory - Autocorrect Vocabulary Candidates
  - Assignment 2 is **due**
  - Assignment 3 is assigned - *Autocorrect and Minimum Edit Distances*

WEEK 5

Introduction to Language Modeling

February 9

- Today we'll begin our journey to understanding LLMs by observing its origins. Dropping the "Large" from the now-ubiquitous term "Large Language Models", we take a look at the foundational principles that extract the relationships defining what it means to model language and how we might generate text.
- Lecturing Topics
  - What is a language model? (Abstractive vs extractive approaches)
  - Overview of Basic Modeling Approaches
  - The N-Gram Model
  - Out of Vocabulary Words and Smoothing
  - Language Model Evaluation
- Application - Autocompleting words and sentences
- Required Keynote Paper - Model Context Protocol
  - An Introduction to MCP - Anthropic's Blog
  - MCP's Central Documentation
  - DeepLearning.Net - Anthropic's on MCP
- Submissions
  - Laboratory 5.1 - N-Grams Processing
  - Laboratory 5.2 - Out of Vocabulary Words
  - Laboratory 5.3 - Building the Language Model
  - Assignment 3 is **due**
  - Assignment 4 is assigned - *Autocomplete with Topical Information*

WEEK 6

Unsupervised NLP - Topic Modeling

February 16

- This week, we will explore David Blei's contributions to the field, a set of concepts that indirectly attack the age-old question of "**what is *k* in the k-means clustering algorithm**. We will review the hierarchical nature of how to model natural language using Bayesian concepts, where our corpora is processed without preserving the order of words. This week's keynote reading is perhaps the most difficult one that you'll read in this class, since it involves a heavy component of probability and statistics.
- Lecturing Topics
  - Parameter Estimation of a Distribution
  - The Dirichlet Distribution and its Attributes
  - Infinite Bayesian in Topic Models
  - Latent Semantic Indexing and Latent Dirichlet Allocation
  - (Collapsed) Gibbs Sampling, and Optimization

- Required Keynote Paper - Latent Dirichlet Allocation
  - David Blei's Lecture
  - Introductory Blog to Topic Modeling
- Submissions
  - Laboratory - Jupyter Notebooks with GPUs
  - Assignment 4 is **due**
  - Assignment 5 is assigned

WEEK 7

Word Modeling with Self-Supervision

Feburary 23

- Perhaps the most influential paper to have come out of the natural language community is the *word2vec* paper that most general machine learning practioners recognize. You'll find elements of its practice in communities from the information retrieval sciences to modern cyber applications to general ML problems. As it pertains to language models, modeling words is often the first stage in any system pipeline that you may design. This week's lecture reviews word models (including word2vec as well as continuous bags of words) and the embeddings / representations that they create.
- Lecturing Topics
  - Embeddings with Continuous Bag of Words
  - Intrinsic and Extrinsice Evaluation of Word Models
  - Word Modeling in Practice
  - The Skip-gram and Negative Sampling
  - From Words to Sentences
- Required Keynote Paper - Distributed Representations of Words
  - Blog - Gentle Introduction to Negative Sampling
- Submissions
  - Laboratory - Word Embeddings with CBOW
  - Laboratory - The Original Word2Vec Code in C (Optional)

WEEK 8

Enjoy Your Spring Break!

March 2

WEEK 9

Introduction to Sequential Modeling

March 9

- In smaller data regimes and for bespoke problems, we'll find that conversations still revolve around the OG state-driven modeling approach - Hidden Markov Models (HMMs). HMMs dominated NLP and speech recognition for half a century and are classic examples of NLP fundamentals. We'll build a strong foundation this week that appreciates the origins of our field, which helps us understand the motivations behind the latest Deep Learning more modern paradigms like Transformers.
- Lecturing Topics
  - Modeling with Hidden Markov Models
  - The Viterbi Algorithm - Initialization, Forward, and Backward Passes
- Application - Parts of Speech Tagging
- Required Keynote Reading - A Survey of LLMs Including ChatGPT and GPT-4
- Required Keynote Reading - Learning Text Similarity with Siamese Recurrent Networks
- Submissions
  - Laboratory - HMMs Text Processing
  - Laboratory - HMMs Numpy PoS Processing

WEEK 10

Recurrence and Neural Networks

March 16

- While newer architectures like transformers now dominate the field of NLP, in its short tenure, Recurrent Neural Networks became workhorses that first demonstrated the power of deep learning for sequential data like text. This lecture builds an appreciation of how modeling language works, how attention and transformers originated, and subsequently the transition to truly deep architectures. Beyond studying the history; it's we'll review the fundamental principles in RNNs that underpin modern NLP.
- Lecturing Topics

        - The Recurrent Neural Network
        - Vanishing and Exploding Gradients
        - Memory Gating - GRUs and LSTMs
        - Accuracy and Evaluation - Perplexity
    - Applications - Named Entity Recognition and Machine Translation
    - Required Keynote Paper Long Short Term Memory Networks
    - Required Keynote Paper - On the Difficulty of Training RNNs
        - Karpathy's Blog on Recurrent Networks
    - Submissions
        - Laboratory - Building Your First RNN
        - Assignment 5 is **due**
        - Assignment 6 is assigned - *Implement Your Own Recurrent Network*

WEEK 11

Attention and the Transformer Model

March 23

- Attention models have been the leap forward that are the fundamental building blocks to modern machine learning today, including the essential ingredients for Large Language Models. We'll go deep into attention layers in neural networks, building our own from scratch.
- Lecturing Topics
    - Introduction to the Attention Modeling
    - The Self-Attention Mechanism
    - The Transformer Modeling Layer
    - Large Scale Attention Modeling
- Required Keynote Reading - Attention is All You Need
- Required Keynote Reading - BERT - Pre-training Bidirectional Transformers
    - BERT Explained - State of the art in NLP, Blog
    - Attention Paper Explained
- Submissions
    - Laboratory - Dot Product Attention
    - Laboratory - Masking in Attention
    - Laboratory - Positional Encoding
    - Assignment 6 is **due**
    - Assignment 7 is assigned - *Attention and Transformer Networks*

WEEK 12

Introduction to Large Language Modeling (LLMs)

March 30

- The next two weeks are devoted to the state of the art in industry, and LLMs in practice, which may have changed in the time that you have started this course! This week, we introduce large language models using the fundamentals that you have learned, from perplexity in system design to transformer neural network layers for pre-training. We'll focus on techniques that large companies (or well-funded ones, at least) use to create *foundation* LLM models, taking training methods from OpenAI, Anthropic, Amazon, and Google.
- Lecturing Topics
    - Large Language Modeling (LLM) in Code
    - Aligning LLMs in the Instruction Following Framework
    - Tuning with Low Resources - LoRA and Quantization
- Required Keynote Reading - Training to Instruct with Human Feedback
- Required Keynote Reading - GPT-4 Technical Report from OpenAI
- Submissions
    - Project Proposals are **due**
    - Laboratory - Serving an LLM
    - Laboratory - Tuning LLMs (Optional)

WEEK 13

Practically Leveraging LLMs and the LLM Lifecycle

April 6

- Last week, we discussed how large companies might train LLMs. In contrast, this week's lecture is most useful for those interested in entering the industry at the mid- to startup levels, where we explore common approaches to optimally *leverage* large language models for your particular applications once the LLM has been created. These techniques

Additionally, we explore the practical aspects of GenAI engineers when product managers ask them to design a system for them. More than the theory, we'll learn about the system itself, devoting time for *when* to focus on certain components of your LLM, and the life cycle of your system design.

- Lecturing Topics
  - Prompt Engineering - Query and Context
  - Retrieval Augmented Generation (RAG)
  - Tuning with Low Resources - LoRA and Quantization
  - Intelligent Agents with Program-Aided LLMs
  - Guidelines and NLP Systems Engineering Diagrams
  - Intelligent Agents with Program-Aided LLMs
  - Multimodal Large Language Models
- Required Keynote Reading - Retrieval Augmented Generation
- Required Keynote Reading - Parameter Efficient Fine-Tuning
- Submissions
  - Laboratory - Instruction Following Tuning
  - Assignment 7 is **due**

WEEK 16

Demonstrations and Poster Sessions

April 13

- Deploy and show off your domain-specific LLM and pitch your startup idea! Review the guidelines at the Final Project Website.

grading criterion

Participation	5%
Reading Group	15%
Labs	25%
LLM Deployment Project	25%
Assignments	30%

grading policies

Labs must be turned in by the following lecture

Late labs / assignments have 3pt deduction each weekday until the following lecture

Late assignments cannot be accepted past the time when the next homework assignment is released

You must notify cs6120-staff@ccs.neu.edu ahead of any absences.

course meeting times

Lectures

- Mon, 4pm-7:20pm
- Room San Jose R1045

Office Hours

- Karl, Thu 8:30-9:30pm, Teams
- Dharun, Fri 12-2pm, 9th Floor
- Swathi, Wed 12-2pm, 9th Floor
- Qunnie, Thu 1-3pm, 10th Floor, Zoom

suggested textbooks

Speech and Language Processing, 3rd Ed. Dan Jurafsky and James Martin, 2024

A Comprehensive Overview of Large Language Models, Naveed et. al., 2024