Tentative Syllabus

Introduction to Speech and Language Processing (CS 5319)
Topics in Data Science (CS 4364)

Spring 2022

Tuesdays and Thursdays 10:30 - 11:50 in Business room 302

Instructor:	Nigel Ward
Office:	CCS 3.0408
Phone:	747-6827
E-mail	nigel@utep.edu
Office Hours:	TBD, also by appointment, and usually when the door is open

Teaching Assistant (for 4364): Jonatan Contreras <jmcontreras2@miners.utep.edu>

Speech and language processing has recently seen tremendous advances, and several core technologies are now mature. There are well-designed systems used by millions of people every day, and readily usable APIs available for both data scientists and developers of user-facing systems. At the same time, the range of viable applications is still quite limited, due to numerous challenging open problems.

This class will provide a survey and sampling of the techniques and issues in speech and language processing. Students will design, implement, and evaluate a project applying these techniques to a problem of their choosing.

Learning Outcomes
> **3a.** Given a well-formulated problem requiring natural language processing, design and implement a solution, by: goal setting; performance metric design; decomposition of the task into modules; selection of components, tools, and resources; implementation; and performance analysis.
> **2a**. Given a user need or a business need related to natural language, identify possible technical solutions, and estimate their feasibility and likely cost.
> **2b**. More rapidly develop software, especially using skills in scripting and in the customization and combination of existing tools.
> **2b.** Comfortably use basic machine learning concepts and techniques.
> **1a.** Apply knowledge of language and of English to improve everyday written and spoken communication, including computer-mediated communication, personally and for groups, organizations, and society.

Main Topics

> **Language**: Basic properties of human language; symbolic, feature-based, vector-space and other representations of each level (acoustic, phonetic, prosodic, morphological, syntactic, semantic, pragmatic), with attention to differences across languages, genres, and speakers.

> **Models and Algorithms**: Standard and for-purpose models and algorithms for speech recognition and other core language processing tasks, including techniques for model training.

Tools and Technologies for corpus wrangling and analysis of text and speech data in support of discovering knowledge from data, including sentiment analysis, filtering, and various classification tasks

Systems: The design and development of systems for search, question-answering, conversational interaction, user state identification, information extraction, and other applications.

Prerequisites: Linear Algebra (Math 3323 or equivalent), Probability and Statistics (Stat 3320 or equivalent), good programming skills (CS 3331 or equivalent), and at least nascent problem-solving and systems-integration skills. Graduate students from other departments may receive prerequisite waivers; see the instructor for permission.

Also helpful will be knowledge of Python, of machine learning techniques, and of basic linguistics concepts, but these are not required.

Format Lectures, student presentations, discussions, in-class design exercises, lab time, project activities, project presentations, guest speakers (tentatively Drs. Thomas and Mueller). There is no remote option.

Textbook *Speech and Language Processing*, Daniel Jurafsky and James H. Martin, 3nd edition, Prentice-Hall, 2021, available at https://web.stanford.edu/~jurafsky/slp3/ . We will be skipping back and forth in the book as we follow the topics listed above. Students wishing to get ahead may read the first few sections of each of Chapters 2-4, 6, 12, 15-16, 18-19, and 25-28.

Course Website http://www.cs.utep.edu/nigel/slp/

Assignments There will be a number of structured assignments, designed to reinforce knowledge and hone skills. Most assignments will be done in teams. Writing quality is important, and rework may be required if not up to standard. Graduate students will have two additional assignments. Late assignments will receive at most 90% credit, less when the solution has been discussed in class, decreasing by 10% per day late.

Grading Approximate point values:
190 assignments, including presentations
  80 project
230 midterms and final exam
  90 participation and quizzes
590 total

Grading will be on a points-earned basis (points above zero), rather than a points-off basis (points below expectation), and everything will be challenging. However letter grades will be assigned appropriately; in the past, the A/B break has been around 80% and the B/C break around 70%.

Conduct, etc. Students are expected to be punctual, and to follow the spirit and letter of the UTEP Standards of Student Conduct and Academic Integrity policy https://www.utep.edu/student-affairs/osccr/student-conduct/academic-integrity.html . In particular, cooperation among students and among teams is encouraged, but not to the extent that it interferes with each individual's understanding or with learning-by-doing. Help given to and received from other students and sources should be noted in the assignment write-up.

If you have or suspect a disability and need accommodation please contact CASS at 747-5148 or at cass@utep.edu or visit Room 106 Union East Building.

Tests will generally be closed-book, except that one page of hand-written notes may be used for the first test, two for the second test, and three for the final. If you leave the classroom for any reason, your test will be graded on only what you did up until that time. No make-up exams or assignments will be given except under the conditions set forth in the Catalog. Students are free to attend class or not, bearing in mind that absence may annoy other students, interfere with learning, and result in a lower grade.

Important Dates
>January 18: Class begins
>February 15: Test 1 (tentative)
>Sring Break: March 14-18
>March 29: Test 2 (tentative)
>May 12, 10-12:45: Final Exam

Tentative Schedule

A. Introduction (1 day)
   a. Overview of Language Applications
   b. Review of Bayes Law and other Basic Mathematics
   Exercise 1: Observe Language in Use (1hr, 5 pts)
   Exercise 2: Link Analysis (5 pts)

B. Rules, Features, and Classification (4 days)
   a. Rules and Tendencies
   b. Feature Design
   c. Linear Classification and Prediction
   d. Model Evaluation
   Assignment B (1,2,3): Simple Predictions (Surnames) (30 points)

C. Sequences, Context, Ngrams, and Language Modeling (3 days)
   a. Edit Distances
   b. The Noisy Channel Model
   c. Bigrams and Beyond
   d. Sequence-to-Sequence Mapping
   e. Tagging
   f. Language Modeling
   Assignment C (1,2,3): Sequence Modeling (Surnames, again) (20)

D. Pattern Matching and Regular Expressions (1 day)
   a. Regular Expressions in Python
   b. Tokenization
   c. Finite State Morphology
   d. Patterns for Shallow Response Generation
   Assignment D (1,2,3,4): Regular Expressions (Chatbots) (20)

Test 1

E. Representations of Meaning (5 days)
   a. Logic-Based
   b. Entities and Graph-Based Meaning Representations
   c. Bag-of-Words
   d. Vector-Space Similarity
   e. Word Embeddings and Context Vectors
   f. Uses in Information Retrieval and Lexical Disambiguation
   Assignment E: Information Retrieval (15)
   Assignment F (1,2,3): Word Embeddings (15)

F. Grammatical Structure (2 days)
   a. Dependencies
   b. Constituency, Context-Free Grammars, and Syntactic Ambiguity
   c. Chunks
      Assignment G: Sentiment Analysis (10)
      Exercise 3: English Grammar (5)

G. Sound, Phonetics, and Prosody (5 days)
   a. Articulatory Phonetics and Phonemes
   b. Acoustic Phonetics and Spectral Representations
   c. The Noisy Channel Model, again
   d. Speech Recognition: Search-based and Transducer-based Models

                                                      Test 2
   e. Speech Recognition Issues
   f. Speech Synthesis
   g. Inferring Speaker Straits and Traits
   h. Call-Center Analytics
      Exercise 4: Phonetic Observations (5)
      Exercise 5: Explorations in Speech Recognition (10)

H. Dialog Structure and Dialog Flow (3 days)
   a. Finite-state Dialog Management
   b. Question-Answering, Retrieval-Based Dialog, and Chatbots
   c. Endpointing and Turn Taking
   d. Pragmatics, Dialog Acts, and User Intentions
   e. Language and Social Impact, across genres and media
   f. Response Tuning and Natural Language Generation
   g. API Interactions
   h. Advantages and Disadvantages of Natural Language Interfaces
      Exercise 6: Dialog Flow Authoring
      Exercise 7: Dialog States and Prediction
      Assignment K: A Minimal Spoken Dialog System (10)

I. Other Applications (1 day)
   a. Spam Filtering
   b. Plagiarism Detection
   c. Machine Translation
   d. Information Extraction
   e. Language Proficiency Assessment
   f. Tutoring and Training Systems
   g. Summarization
   h. Search, Collaborative Filtering, and Recommendations
      Exercise 8: The Business Landscape (5)
      Exercise 9: Ethical Issues (5)

Other Assignments
   Assignment P: Final Project (80)
   Assignment X: Present a Research Paper (graduate students only) (30)
   Assignment Y: Research-Project Mini-Proposal (graduate students only) (20)
   Exercise 10: A Question for the final exam (5)