August 12, 2020

Tentative Syllabus

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

Fall 2020

Tuesdays 10:30 - 11:50 in CCSB 1.0702 and Thursdays 10:30 – 11:50 in Blackboard.
Room subject to change. An online alternative for Tuesdays will be available.

Instructor:	Nigel Ward
Office:	CCS 3.0408
Phone:	747-6827
E-mail	nigel@utep.edu
Office Hours:	Mondays and Wednesday 1-2 in Blackboard Collaborate, and by appointment

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.
2c. 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 (acoustics, phonetics, prosody, morphology, syntax, meaning, pragmatics), 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), 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, knowledge 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. Mueller and Novick).

Face-to-face every Tuesday until Thanksgiving, and otherwise online, but subject to change. Fully online attendance will also be an option, and is required if you feel sick, etc., as detailed below.

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/

Grading and approximate point values

assignments, including presentations (190)
project (80)
midterms and final exam (230)
participation and quizzes (90)
total (590)

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. Letter grades will be assigned appropriately; in the past, the A/B break has been around 80% and the B/C break around 70%. The gradebook in Blackboard is not reliable; actual grade-to-date information will be provided periodically, and at any time upon request.

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.

Assignments will be generally due at 10:28 in Blackboard or 10:30 if you choose to submit hardcopy. Late assignments will receive at most 90% credit, less when the solution has been discussed in class, decreasing by 10% per day late.

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.

Tests The format of tests remains to be determined. Tests may require Respondus Lockdown Browser. The final exam may be face-to-face. Tests will most likely 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. For in-classroom tests, if you leave the room 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.

Participation On Thursdays, class will be synchronous: everyone will be expected to attend in Blackboard Collaborate. On Tuesdays, online students should initially expect to attend at the scheduled time in Blackboard Collaborate, but this may change.

Participation credit will be based on live participation, either in-person or through Blackboard, and on postings on the discussion boards. Visual attention and feedback counts towards participation, so during online sessions keep your webcam on as much as possible. Fully online students will have additional postings assigned and in general should post more. Postings that are especially helpful to the class, for example to General Questions pointing out the need for clarification on an assignment or answering another students' questions, are greatly appreciated and may be rewarded.

UTEP-General Information

Academic Integrity Students will 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 . Suspected violations will be reported.

Disabilities 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.

Blackboard Course content is delivered via the Internet through the Blackboard learning management system. Ensure your UTEP e-mail account is working and that you have access to the Web and a stable web browser. Google Chrome and Mozilla Firefox are the best browsers for Blackboard; other browsers may cause complications. When having technical difficulties, update your browser, clear your cache, or try switching to another browser. The UTEP Helpdesk can help if you have problems.

Equipment For online sessions, besides your computer/laptop you will need a webcam with microphone. You will also need a scanner or camera in order to upload images of hand-drawn designs and solutions.

Copyright Course materials, recordings, and Blackboard postings are private, and should not be reposted to any publicly accessible website etc.

Covid-19 You must STAY AT HOME and REPORT if you (1) have been diagnosed with COVID-19, (2) are experiencing COVID-19 symptoms, or (3) have had recent contact with a person who has received a positive coronavirus test. Reports should be made at screening.utep.edu. If you know of anyone who should report any of these three criteria, you should encourage them to report. If the individual cannot report, you can report on their behalf by sending an email to COVIDaction@utep.edu.

For each day that you attend campus—for any reason—you must complete the questions on the UTEP screening website (screening.utep.edu) prior to arriving on campus. The website will verify if you are permitted to come to campus. Under no circumstances should anyone come to class when feeling ill or exhibiting any of the known COVID-19 symptoms.

Wear face coverings when in common areas of campus or when others are present. You must wear a face covering over your nose and mouth at all times in this class. If you choose not to wear a face covering, you may not enter the classroom. If you remove your face covering, you will be asked to put it on or leave the classroom. Students who refuse to wear a face covering and follow preventive COVID-19 guidelines will be dismissed from the class and will be subject to disciplinary action according to Section 1.2.3 *Health and Safety* and Section 1.2.2.5 *Disruptions* in the UTEP Handbook of Operating Procedures. If unable to wear a face covering (e.g., medical reasons), the best course of action is to take the online option.

• You are encouraged to complete Covid training at https://covidtraining.questionpro.com/ . • Contact instructor if temporary accommodations due to COVID-19 are needed (i.e., due to positive COVID-19 test, symptoms, or exposure). • Maintain 6 feet of separation at all times, including when talking with other students. • Follow signage indicating specific entry and exit doors and pathways. • Do not cluster in groups and keep hallways open. • Wash hands and/or apply hand sanitizer prior to entering classroom and after leaving a classroom. Do not touch face until after hands are washed/sanitized. • Use an alcohol wipe, provided outside of classrooms, to sanitize the desk, chair, or table. • Follow faculty protocols for leaving and re-entering the classroom.

Important Dates

August 25: Class begins
September 24: Test 1 (tentative)
October 27: Test 2 (tentative)
November 26: Thanksgiving (no class)
December 10: Final Exam, 10:00-12:45

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. Information Retrieval
g. 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, Training
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)