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

Fall 2023

Tuesdays and Thursdays 3:00 – 4:20 in Education Room 112

Instructor: Nigel Ward
Office: CCS 3.0408
Phone: 747-6827
E-mail nigel@utep.edu
Office Hours: Wed 3:45 – 4:45, Fri 1:30 – 2:30, usually when the door is open, 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 what they learn 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 choice; decomposition of the task into stages and modules; selection of models, 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 configuring and combining 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.
1b. Explain how speech and language processing relates to AI and data mining methods and needs.

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, domains, genres, and speakers.

Representations, Models and Algorithms: Standard and for-purpose, for various core language processing tasks, including techniques for model training.
**Tools and Technologies**: Ways to wrangle corpora and analyze 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. There is no remote option.

**Textbook**  

**Code Resources**  
Well-known toolsets include NLTK, SpaCy, and gensim. At github, ekochmar and vajjala have useful examples in Jupyter.

**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 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
August 29  Class begins
September 28  Test 1 (tentative)
November 2  Test 2 (tentative)
November 23  Thanksgiving
December 14, 4:00-6:45  Final Exam

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. Text Classification with a Linear Model
   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
   g. Generative Modeling and Large Language Models
      *Assignment C (1,2,3): Sequence Modeling (Surnames, again) (20)*
      *Assignment X: Playing with Large Language Models (XX)*

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

Test 1  (5 days)
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. Lexical Disambiguation  
f. Word Embeddings and Context Vectors  
g. Information Retrieval  
h. Large Language Model Training
   Assignment E: Information Retrieval (15)  
   Assignment F: Using Large Language Models or Word Embeddings (20)

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

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. Response Tuning and Natural Language Generation  
f. Advantages and Disadvantages of Natural Language Interfaces
   Exercise 6: Dialog States and Flow Authoring (5)  
   Assignment K: A Minimal Spoken Dialog System (10)

I. Other Applications (1 day)  
a. Search  
b. Spelling Correction  
c. Spam Filtering  
d. Machine Translation  
e. Information Extraction (Entity Extraction and Linking)  
f. Language Proficiency Assessment  
g. Tutoring and Training  
h. Summarization
   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)
Readings

Jurafsky and Martin, *Speech and Language Processing*, 3rd edition, January 2022 version
- Chapter 2, except byte-pair encoding (Regular Expressions, etc.)
- Chapter 3 through 3.4 (Language Modeling)
- Chapter 4, Sections 4.7, 4.8 (Sentiment)
- Chapter 6 through 6.6, 6.8 - 6.12 (Vector Space Models, Embeddings, Information Retrieval)
- Chapter 8, through 8.3 (Part-of-Speech Tagging)
- Chapter 10, through 10.1, historical notes (Machine Translation)
- Chapter 11 introduction (Pretrained Models)
- Chapter 12 through 12.4.1 (Constituency Structures)
- Chapter 13 through 13.1 (Parsing)
- Chapter 14 through 14.1 (Dependency Structures)
- Chapter 15 through 15.3 (Logical Representations of Sentence Meaning)
- Chapter 17 through 17.1.2, 18.2 through 18.2.1, 18.3, 18.5 (Information Extraction)
- Chapter 18 through 18.4 (Word Senses)
- Chapter 20 through 20.2 (Lexicons for Sentiment)
- Chapter 23 through 23.1, 23.2 through 23.1.2, 23.3 (Question Answering)
- Chapter 24 (Dialog Systems and Chatbots)
- Chapter 25 (Acoustic Phonetics)
- Chapter 26 through 26.2, 26.6-26.7 (Speech Recognition, Text to Speech)

Manning, Raghavan, Schutze: *Introduction to Information Retrieval*
- Chapter 19, pp 421-434
- Chapter 21 through 21.2.2

Giangola, Balogh, and Cohen, *Voice User Interface Design*
- Chapter 3 (Overview of the Methodology)

Brin and Page: *The Anatomy of a Large-Scale Hypertextual Web Search Engine*

Sarikaya: *The Technology behind Personal Digital Assistants*

Links for most are on the course homepage.