

CS 5761 : Introduction to Natural Language Processing (4)**Catalog Description:**

Techniques for creating computer programs that analyze, generate, and understand natural human language. Topics include syntactic analysis, semantic interpretation, and discourse processing. Applications selected from speech recognition, conversational agents, machine translation, and language generation. Substantial programming project required.

Textbook: *Speech and Language Processing : an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* by Daniel Jurafsky and James H. Martin. Prentice-Hall. (2nd edition) 2008.

References: *Learning Perl*, by Randal L. Schwartz, Tom Phoenix, and brian d foy.. O'Reilly Publishers. (4^d edition) 2005.

Course Goals: The goal of this course is to provide students with a sufficient foundation in both knowledge and corpus based natural language processing so that they can complete a significant individual research project by the end of the semester. Students are exposed to a range of natural language processing problems that lend themselves to solutions based on statistical methods, artificial intelligence, and formal languages/automata theory. As such students are expected to understand the characteristics of rule-based versus empirical methods, and be able to appropriately choose between them.

Prerequisites by Course & Topic

CS 3512 – Computer Science Theory – formal languages and automata theory and (via 2511 prerequisite of 3512), software design and implementation

Major Topics Covered in the Course

History of NLP
Regular Expressions and Automata
Finite State Transducers
Probabilistic Models of Spelling
N-grams
Part of Speech Tagging
Context Free Grammars for English
Representing Meaning
Semantic Analysis
Lexical Semantics
Word Sense Disambiguation
Conversational Agents

Class/Laboratory Schedule: Lecture: 3 hours per week, Laboratory: 1

Course Outcomes

1. Learn computational techniques suitable for analyzing, generating, and understanding natural human language.
 - a. Apply classical techniques from computer science to real-world problems in natural language processing.
 - b. Analyze the limitations of these classical techniques, and be able to appropriately augment or replace them with techniques from Statistics and Artificial Intelligence.
 - c. Understand the historical/intellectual origins of Natural Language Processing.
2. Develop analysis, design, evaluation, and dissemination skills appropriate for conducting research.
 - a. Identify a research topic of suitable scope and complexity for the available time.
 - b. Prepare a written project proposal, which includes a detailed plan for how the project will be evaluated and assessed.
 - c. Carry out the proposed work, including the evaluation.

- d. Prepare a written summary of their findings, focusing on the insights gained about the problem under investigation.
- 3. Develop programming skills in a scripting language, and making extensive use of regular expressions.
 - a. Implement significant projects in the Perl programming language.
 - b. Understand the design and optimal use of a sophisticated regular expression matching engine.

Relationship to Program Outcomes

The prerequisite of CS 5761 is computer science theory. This course contributes to meeting the following program outcomes:

- 2. *Students can design, develop, and analyze significant software systems.*
 Students gain experience in analyzing the limitations of applying classical techniques to NLP problems and learn to replace them with techniques from AI and statistics. They must prepare detailed written project proposals including how the project will be evaluated and assessed, and also a final written report that describes the outcomes of their experiments. All three course outcomes map to this program outcome.
- 3. *Students understand data structures and related algorithms, programming languages, and the fundamentals of computer organization and architecture.*
 Students use regular expressions and finite state automata to represent language processing problems. They make extensive use of multi-level hash tables. They learn Perl and implement significant projects using it. All three course outcomes map to this program outcome.
- 4. *Students can apply computer science principles and practices to a variety of problems.*
 Students have the opportunity to develop software that analyzes, generates and understands natural language. This requires that they apply techniques from automata theory (when using regular expressions and finite state tools), statistics and Artificial Intelligence, and NLP (when carrying out basic tasks such as parsing and part-of-speech tagging which are prerequisites to most projects). All three course outcomes map to this program outcome.
- 9. *Students understand the scientific method and can apply this mode of inquiry in a laboratory setting.*
 The research project required in this class asks the student to identify an interesting and significant problem, and to formulate a hypothesis as to how it could be approached effectively. The purpose of the project is not to simply implement a solution, but rather to evaluate their hypotheses and thereby gain insight into the problem they are attacking. All three course outcomes map to this program outcome.

Estimate CSAB Category Content

	CORE	ADVANCED		CORE	ADVANCED
Data Structures		1	Computer Organization and Architecture Concept of Programming Languages		
Algorithms		1			1
Software Design		1			

Coordinator/Prepared by: T. Pedersen (March 2009)