UNIVERSITY OF NEBRASKA AT OMAHA
COURSE SYLLABUS

Department and Course Number	CSCI 3450
Course Title	Natural Language Processing
Course Coordinator	Yuliya Lierler
Total Credits	3
Repeat for Credit?	No
Date of Last Revision	September 18, 2015

1.0 Course Description Information

1.1 Catalog description:

The course will provide overview of the topics in natural language processing such as word and sentence tokenization, syntactic parsing, semantic role labeling, text classification. We will discuss fundamental algorithms and mathematical models for processing natural language, and how these can be used to solve practical problems. We will touch on such applications of natural language processing technology as information extraction and sentiment analysis.

1.2 Prerequisites of the course:

CSCI 2030 Mathematical Foundations of Computer Science

Co-requisites of the course:

CSCI 3320 Data Structures

Students planning to enroll in this course should be comfortable with scripting (preferably Python as this is the language extensively utilized in natural language processing tools including NLTK – natural language tool kit – that will be used in a course).

1.3 Suggested preparatory courses:

1.4 Overview of content and purpose of the course:

This course will span topics in natural language processing (NLP). Advances in NLP aim to address the problem of knowledge extraction from natural language content – a task that is of critical importance of late in a world with so much information easily accessible for content mining. The goal of the course is to equip students with an understanding of fundamental problems and solutions in NLP, knowledge about off-the-shelf NLP tools, hands-on experience of gluing these tools together while building an in-house NLP system, and the desire to tackle key issues in the domain. We will specifically focus on the building blocks of NLP applications, including word and sentence tokenization, spelling correction, syntactic parsing, lexical semantics, text classification, sentiment analysis and information extraction

1.5 Unusual circumstances of the course. None

2.0 Course Justification Information

2.1 Anticipated audience / demand:

This course is intended for undergraduate students in computer science and related disciplines interested in studying computational methods in natural language processing and data mining.

2.2 Indicate how often this course will be offered and the anticipated enrollment:

This course will be offered every spring semester. The anticipated enrollment is 25 students per semester.

2.3 If it is a significant change to an existing course, please explain why it is needed:

CSCI 4460 is a current course titled "Natural Language Understanding" and offered as a senior class. This is a proposal to narrow the scope of the course to accommodate junior students in IS&T. The goal of this change is to equip students with computational methods of NLP earlier in their education. The NLP methods provide a strong foundation for projects in data mining and information extraction, areas that cut across a variety of computer science and information technology subjects. In their senior year students will have a chance to connect their knowledge in NLP methods to their capstone projects as well as other courses in the college.

3.0 List of performance objectives stated in learning outcomes in a student's perspective:

Upon completion of this course, the student should:

3.1 Have a clear understanding of building blocks of key components of natural language processing systems: part of speech tagging, syntactic and semantic analysis

3.2 Understand basic underlying theory from probability, statistics, and machine learning that are crucial for the field of NLP

3.3 Know about some NLP of-the-shelf tools and resources available such as NLTK, Wordnet, and Framenet

3.4 Be able to build NLP applications such as text classifier (spam vs non-spam) or sentiment analyzer

3.5 Understand the techniques for assessing the performance and results produced by NLP tools

4.0 Content and Organization Information

4.1 List the major topics central to this course:

4.1.1 Introduction to Natural Language Processing and its Applications (Overview of fundamental tasks): 3 contact hours

4.1.1.1 Word Segmentation

4.1.1.2 Morphological Analysis

4.1.1.3 Part of Speech Tagging

4.1.1.4 Phrase Chunking

4.1.1.5 Syntactic Parsing

4.1.1.6 Semantic Role Labeling
4.1.1.7 Semantic Parsing
4.1.1.8 Information Extraction
4.1.1.9 Sentiment Analysis
4.1.1.10 Machine Translation

4.1.2 Basic Language Processing Parsing: 3 contact hours
4.1.2.1 Word Tokenization
4.1.2.2 Word Stemming
4.1.2.3 Sentence segmentation

4.1.3 Introduction to Basics of Machine Learning: 4.5 contact hours
4.1.3.1 Basics of Machine Learning
4.1.3.2 Probabilistic Learning
4.1.3.3 Naive Bayes Learning
4.1.3.4 Smoothening
4.1.3.5 Evaluation
4.1.3.6 WEKA: software "workbench" of machine learning techniques

4.1.4 Language Models: 4.5 contact hours
4.1.4.1 Corpora: PennTreebank, Brown
4.1.4.2 Corpora and Word Counting
4.1.4.3 N-Grams
4.1.4.4 Applications of N-Grams

4.1.5 Part-of-Speech Tagging: 3 contact hours
4.1.5.1 Tagsets and English
4.1.5.2 Rule-based Part of Speech Tagging
4.1.5.3 Part of Speech Tagging as a Sequence Classification Problem
4.1.5.4 Hidden Markov Models Part of Speech Tagging

4.1.6 Syntactic Parsing: 4.5 contact hours
4.1.6.1 Phrase Chunking
4.1.6.2 Syntactic Parsing
4.1.6.3 Top Down, Bottom Up Parsing
4.1.6.4 CYK Chart Parsing

4.1.7 Natural Language Tool Kit – NLTK: 3 contact hours
4.1.7.1 Introduction to NLTK Project
4.1.7.2 NLTK for Part of Speech Tagging

4.1.7.3 NLTK for Parsing

4.1.8 Computational Semantics: 4.5 hours

4.1.8.1 Semantic Representation and Logic forms

4.1.8.2 Lexical semantics

- Wordnet
- Framenet

4.1.8.3 Principle of Compositionality

4.1.8.4 Inference

4.1.8.5 Machine Learning Approach

4.1.9 Semantic Role Labeling: 3 hours

4.1.9.1 Semantic roles

4.1.9.2 Syntactic cues and restrictions

4.1.10 NLP Applications: 6 hours

4.1.10.1 Information Extraction

- Semantic role labeling as information extraction
- World wide web and other applications
- Evaluation methods

4.1.10.2 Sentiment Analysis

- Subjectivity and objectivity in sentiment analysis
- Feature-based sentiment analysis
- Algorithmic methods
- Evaluation methods

The contact hours account for 13 weeks leaving 1 more week in a semester for in-class presentations and review session.

5.0 Teaching Methodology Information

5.1 Methods:

The course will be in lecture form.

5.2 Student role:

The students will be actively involved in the class through programming, homework, and reading assignments. The students will have to write a technical report on their final project as well as present a demo on their work.

6.0 Evaluation Information

6.1 Describe the typical types of student projects that will be the basis for evaluating student performance:

Evaluation of student performance will be based on programming, homework, and reading assignments, a project presentation/report, and a final. Class participation will be taken into account.

Sample Course Project: Students will implement a natural language system for the task of resolving prepositional phrase attachment problem in syntactic parsing that will take into account information present in lexicons such as Verbnet and Wordnet.

Programming and homework assignments will be designed as building blocks of the core of the system for prepositional phrase attachment to ensure a steady progress towards the final project.

Homework and reading assignments will be based on key technical topics discussed in class. These will follow the textbook of the course – Speech and Language Processing, 2nd edition by Jurafsky and Martin – and will supplement material presented in class.

Students will submit one extended written report describing the technical details of their course project.

Final will assess students' understanding of NLP topics covered in the course.

6.2 Describe the typical basis for determining the final grade (e.g. weighting of various student projects):

25% Assignments
25% Final Project
20% Final Exam
20% Reports and Presentations
10% Class Participation

6.3 Grading type:

97 – 100	*A+*	*77 – 79*	*C+*
94 – 96	*A*	*70 – 76*	*C*
90 – 93	*A–*	*70 – 73*	*C–*
87 – 89	*B+*	*67 – 69*	*D+*
84 – 86	*B*	*64 – 66*	*D*
80 – 83	*B–*	*60 – 63*	*D–*
		0—59	*F*

7.0 Resource Material Information

7.1 Text book:

Jurafsky, D., & Martin, J. (2009). *Speech and Language Processing* (2nd ed.).

Handouts given in class and technical papers.

7.2 Other student suggested reading materials:

Allen, J. (1995). *Natural Language Understanding.* The Benjamin/Cummings Publishing Company, Inc.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly.

Perkins, J. (2010). *Python Text Processing with NLTK 2.0 Cookbook.* Packt Publishing

7.3 Bibliography

Allen, J. (1995). *Natural Language Understanding.* The Benjamin/Cummings Publishing Company, Inc.

Blackburn, P., & Bos, J. (2005). *Representation and Inference for Natural Language: A First Course in Computational Semantics.* CSLI Publications.

Damerau, F., & Indurkhya, N. (2010). *Handbook of Natural Language Processing.* (2nd ed.). CRC Press.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition.* MIT Press.

Frank, E., & Hall, M., & Witten, I. (2011). *Data Mining Practical Machibe Learning Tools and Techniques.* (3rd ed.). Elsevier.

Jurafsky, D., & Martin, J. (2009). *Speech and Language Processing* (2nd ed.).

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly.

Kamp, H,, & Reyle, U. (1993). *From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory* (Studies in Linguistics and Philosophy). Springer.

Perkins, J. (2010). *Python Text Processing with NLTK 2.0 Cookbook.* Packt Publishing.

Manning, C., & Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press.

Norvig, P., & Russel, S. (2010). *Artificial Intelligence A Modern Approach.* (3rd ed.). Pearson.

8.0 Computer Science Accreditation Board (CSAB) Category Content (class time in hours):

CSAB Category	*Core*	*Advanced*
Data structures	25	10
Computer organization and architecture	0	0
Algorithms and software design	35	15
Concepts of programming languages	10	5

9.0 Oral and Written Communications:

In presentation requirement of the class (presentation on a course project), students will need to use effective public speaking skills. The presentation must include an effective introduction, solid body with main points and supporting materials, and a fitting conclusion. Students may/should use slides/handouts to assist them in the presentation. Students will be also required to produce a written report. It will summarize their findings on their final project.

10.0 Social and Ethical Issues:

None.

11.0 Theoretical content:

Please list the types of theoretical material covered, and estimate the time devoted to such coverage.

12.0 Problem analysis:

Please describe the analysis experiences common to all course sections.
Students experience analysis of problems through class examples and programming assignments.

13.0 Solution design:

Please describe the design experiences common to all course sections.
Students will design natural language processing systems based on the NLTK. They will evaluate their applications according to the metrics common in NLP research.

S – Strong relationship
X – Contributing relationship

Course objective	(a) knowledge of discipline	(b) analyze problem, define requirements	(c) design and implement solution	(d) function on a team	(e) ethical issues	(f) communicate effectively	(g) analyze impact of computing	(h) continued professioanl development	(i) Current techniques and tools	(j) apply foundations	(k) apply design and development principles
1	S								S	S	S
2	S								S		S
3	S	S	S	S		S			S	S	S
4	S	S	S	S		S			S	S	S
5	S								S		S