COURSE NUMBER AND NAME: COSC 43xx / Introduction to Natural Language Processing (ugrad)

CREDITS AND CONTACT HOURS: 3 credits / 3 lecture hours per week

INSTRUCTOR: Dr. Mukherjee

REQUIRED READING:

- SPEECH and LANGUAGE PROCESSING, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition, by Daniel Jurafsky and James H. Martin, Prentice Hall, 2008.

RECOMMENDED READING:

1. Foundations of Statistical Natural Language Processing, by Christopher D. Manning and Hinrich Schütze, The MIT Press, 1999

2. Introduction to Information Retrieval by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Cambridge University Press. 2008 (online version available at http://www-nlp.stanford.edu/IR-book/).

3. WDM: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Bing Liu; Springer, 1st Edition.

CATALOG COURSE DESCRIPTION:

Introduction to Natural Language Processing. Topics covered include corpus analysis, word collocations, word embeddings, part-of-speech tagging, syntactic parsing, semantic analysis, sequence labeling algorithms, n-gram language models, parsing and grammar formalisms, text categorization, and neural models for language processing.

COURSE DESCRIPTION:

This is a undergraduate level introductory course to natural language processing (NLP). The course is intended to develop basic foundations in NLP and text mining. The broader goal is to both understand how NLP tasks are carried out in the real world (e.g., noisy text, Web, social media conversations) and also to develop the skillset for building models that carry out core NLP tasks (e.g., Language models, vector semantics, embeddings, part of speech tagging, parsing, question answering, information retrieval, etc.). Throughout the course, large emphasis will be placed on developing core NLP building blocks and on applying NLP techniques to specific real-world applications through hands-on experience. The course is standalone and covers required topics in basic mathematical foundations.

PREREQUISITES:

Required prerequisite:

Data Structures (COSC 2436)

Recommended prerequisite:

Data Science I (COSC 3337) or Data Science II (COSC 4337) or Artificial Intelligence (COSC 4368)

COURSE ELECTION: Undergraduate /Elective

COURSE GOALS:

In this course, students will study algorithms and techniques for developing computational models for analyzing, understanding, and generating human language, including parsing techniques, semantic representations, discourse analysis, and statistical and corpus-based methods for text processing and knowledge acquisition. By the end of the course students will have a good understanding of and appreciation for core natural language processing tasks and the general issues related to syntax, semantics, and pragmatics of language. Throughout the course students will be exposed to applications that can benefit from automatic language processing and the state-of-the-art-techniques involved in these applications. They will also learn how to build NLP based applications and tools in the real-world setting.

LEARNING OUTCOMES:

1. Understanding challenges of the field, fundamental methods and key techniques used, and identifying applications of NLP techniques in relevant domains.
2. Ability to leverage existing components of the NLP pipeline to build new tools.
3. Ability to implement core NLP tasks and also develop novel methods (upon building over existing tools) to build real-world NLP applications.
4. Ability to complete practical projects and familiarity with key concepts to understand upcoming papers in the field.

TENTATIVE LIST OF DISCUSSION/LECTURE TOPICS:

- The lexicon, morphology, and word collocations
- N-gram language models
- Vector Semantics
- Word Embeddings
- Markov and Hidden Markov Models for Sequence Tagging
- Part of speech tagging
- Syntactic parsing, grammar formalisms
- Semantic analysis, word sense disambiguation

- Parsing, Role labeling, Coreference Resolution
- Information extraction
- Automatic document summarization
- Information retrieval
- Chatbots and Dialogue Systems
- Sentiment Analysis and Psycholinguistics
- Text clustering and categorization

TENTATIVE GRADING:

The course will be graded based on a combination of homework assignments, projects and exams. Programming homework (60%), Exams (40%)

.