CMSC498 (or new 400 number)
Title: Introduction to Natural Language Processing

Credits: 3

Description: This course will introduce fundamental concepts and techniques for automatically processing and generating natural language with computers. We will study the machine learning techniques, models, and algorithms that enable computers to deal with the ambiguity and implicit structure of natural language. We will apply these techniques in a series of assignments designed to address a core application such as question answering or machine translation.

Grading method: regular

List of prerequisites and/or course restrictions: Minimum grade of C- in CMSC320, CMSC330, and CMSC351; and 1 course with a minimum grade of C- from (MATH240, MATH461); and permission of CMNS-Computer Science department.

First term that the course will be offered: Fall 2018.

Learning Goals
- Acquire the fundamental linguistic concepts that are relevant to automatic processing of language. Assessed in the homeworks/quizzes and exams.
- Analyze and understand state-of-the-art algorithms and machine learning techniques for reasoning about language data. Assessed in the exams, homeworks/quizzes and projects.
- Implement state-of-the-art machine learning algorithms for reasoning about language data. Assessed in the projects.
- Adapt and apply state-of-the-art language technology to new problems and settings. Assessed in the projects.

Course Content
The course will cover fundamental linguistic concepts and machine learning techniques for Natural Language Processing, using a core application (e.g., machine translation or question answering) as a running theme:
- Words and their meaning
  - word sense disambiguation via supervised classification
  - semantic relations and their detection via word embeddings (word association metrics, dimensionality reduction, word2vec)
- Language modeling
  - n-gram and neural language models (feedforward neural networks, recurrent neural networks)
- Language generation
  - Neural sequence-to-sequence models
- Attention mechanism
- Beam search
- Structured prediction
  - Sequence labeling with the structured perceptron, Viterbi algorithm

There will be 3 independent projects that explore complementary building blocks of a single application. For instance, for Machine Translation
  - Project 1: word translation (implement multiclass classification)
  - Project 2: organizing word translations into a sentence (implement language models to score translation hypotheses)
  - Project 3: translation evaluation (implement a simple metric based on string-matching and improve it based on various techniques learned in the course)

Programming projects will be in python.

**Grading scheme**
- 40% 3-programming projects
- 25% ELMS-based homework and quizzes
- 20% Midterm exams (2)
- 15% Final exam

**Readings**
- Primary textbook: Speech and Language Processing, Jurafsky & Martin, *3rd edition draft* freely available online
- Additional readings from the Course In Machine Learning by Hal Daume III (*freely available online*), as well as online research papers and tutorials.