CMSC470
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; 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.

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

Projects will ask students to implement and evaluate building blocks for a core NLP application. For instance, for Machine Translation

- word translation (implement and evaluate multiclass classification)
- sequence translation (implement and evaluate sequence-to-sequence models)
- 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, and will use machine learning libraries such as scikit-learn and pytorch.

Grading scheme

- 30% 3-programming projects
- 30% ELMS-based homework and quizzes
- 40% Exams (2 midterms, 1 final)

Readings

All readings will be selected from books and papers freely available to UMD students.
Primary textbook: Speech and Language Processing, by Dan Jurafsky & James H. Martin, 3rd edition draft
Other references:

- A Course In Machine Learning by Hal Daume III
- Natural Language Processing by Jacob Eisenstein

Experience with Fall 2018 offering

- First offering this semester: http://www.cs.umd.edu/class/fall2018/cmsc470/
- Even in a small class of motivated students, there is a wide spectrum of machine learning backgrounds and levels of mathematical maturity. We are paring down the content initially planned to have more time for examples and exercises for the students who need them. At the same time, we have started incorporating short spotlights on recent research papers to keep more advanced students engaged.
- Tackling real applications in projects is computationally intensive, so current projects are based on toy tasks. We will look into securing additional computing resources in the future so that students can work on more realistic projects.