CSC5035Z (2022): Natural Language Processing

Lecturer: Dr. Jan Buys (jbuys@cs.uct.ac.za)
TA: Francois Meyer (francoisrmeyer@gmail.com)

Course Aims: This course aims to introduce you to the most important applications and models in Natural Language Processing. You will learn how to design and build data-driven systems for classifying, labelling, and generating text. You will be exposed to basic terminology, evaluation techniques, fundamental mathematical/statistical models and algorithms, optimizations and trade-offs, application areas and contemporary topics. You will learn how to formalise natural language processing problems, process datasets, and choose appropriate machine learning models and algorithms to use in your own applications.

Course content: Text processing (tokenization, lemmatization and text normalization). Naive Bayes and logistic regression for text classification (e.g. sentiment analysis). Word vectors and vector semantics. Language models (n-grams and feed-forward neural networks). Sequence labelling with hidden Markov Models (for Parts-of-Speech tagging and Named Entity Recognition). Syntactic dependency parsing. Recurrent neural networks for sequence processing. Machine translation with encoder-decoder neural networks. Transformers and contextual embeddings. Information extraction and question answering.

Prerequisites: Basic calculus, linear algebra, and probability theory. Basic machine learning knowledge is recommended (supervised learning and classification). No prior knowledge about linguistics is required.

The course will run from 18 July to 2 September. The schedule is given below; minor changes may be made during the course.

Lectures:

- First week: Monday, Tuesday, Wednesday, Thursday 11:00-12:00 in CSC302.
- Remaining weeks: Monday, Tuesday, Wednesday 11:00-12:00 in M304.

Assessments:

- Assignment 1: Part-of-Speech Tagging with Hidden Markov Models
- Assignment 2: Language Modelling with Feedforward Neural Networks
- Assignment 3 (for Masters students): Review a recent NLP paper
- Exam: 12-hour take-home (in the week of 29 August)

Marks breakdown:

- Honours students: Assignment 1 (25%), Assignment 2 (25%), Exam (50%)
- Masters students: Assignment 1 (20%), Assignment 2 (20%), Assignment 3 (10%), Exam (50%)

Honours students may optionally do assignment 3 as well, in which case their assignment mark will be the maximum between the allocation including or excluding assignment 3.

Course material: Slides and other resources will be provided with each lecture. Most of the content is covered in:

- Speech and Language Processing (3rd ed. draft). Dan Jurafsky and James H. Martin. Available at: https://web.stanford.edu/~jurafsky/slp3/

Other textbooks that may be useful:
- Neural Network Methods for Natural Language Processing. Yoav Goldberg
- Natural Language Processing. Jacob Eisenstein

Schedule

Week	Monday	Tuesday	Wednesday	
18-07	Introduction	Text processing	Naïve Bayes text classification	(Thursday) Naïve Bayes
25-07	n-gram language models	Hidden Markov Models	Hidden Markov Models	
1-08	Logistic regression for classification	Logistic regression for classification	Feedforward Neural Networks	Assignment 1 due 8-08
8-08	Word vectors	*Public Holiday*	Recurrent Neural Networks	
15-08	Neural Machine Translation; Transformers	Contextualized Representations	Text Generation	Assignment 2 due 24-08
22-08	-	-	-	Assignment 3 due 5-09

Notes revised: 10 August