NATIONAL UNIVERSITY
of Computer & Emerging Sciences, Lahore

Department of Computer Science

CS4063 – Natural Language Processing

Spring 2024

Instructor Name: Maryam Bashir
Email address: maryam.bashir@nu.edu.pk
Office Location/Number: New Building, 022
Office Hours: Monday, Tuesday, Wednesday 11:30 AM to 12:30 PM

Course Information

Program: BSDS **Credit Hours:** 3 **Type:** Elective
Pre-requisites: Linear algebra, Probability and Statistics, Programming in any high level language, Algorithms
Class Meeting Time: BDS-8A: Monday and Wednesday: 10:00 AM to 11:30 AM
Class Venue: NB-207

Course Description:

Natural Language Processing (NLP) addresses fundamental questions at the intersection of human languages and computer science. Understanding complex language utterances is also a crucial part of artificial intelligence. Applications of NLP are everywhere because people communicate almost everything in language: web search, advertisement, emails, customer service, language translation, radiology reports, etc. There are a large variety of underlying tasks and machine learning models behind NLP applications. This course addresses topics like how computers can do useful things with human languages, e.g. filter junk email, extract social networks from the web, and find the main topics in the day's news. This course is also about how computational methods can help linguists explain language phenomena, including automatic discovery of different word senses and phrase structure.

Course Learning Outcomes (CLOs):
At the end of the course students will be able to:
1. Identify techniques for information retrieval, language modeling, smoothing methods
2. Use n grams for language modeling
3. Simulate, apply, or implement classic algorithms for parsing natural language.
4. Simulate, apply, or implement classic and algorithms for named entity recognition, word Sense Disambiguation.
5. Represent words as vectors for a variety of NLP tasks like sentiment analysis, question answering etc.
6. Apply deep learning techniques for sequence modeling

Course Textbook

Jurafsky and Martin, *SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Third Edition*, McGraw Hill. Available Online

Additional references and books related to the course:

1. Chris Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA: May 1999.Steven Bird, Ewan Klein, and Edward Loper.
2. *Natural Language Processing with Python– Analyzing Text with the Natural Language Toolkit*. The first edition of the book, published by O'Reilly, is available at http://nltk.org/book_1ed/
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need**. In Proceedings of *the 31st International Conference on Neural Information Processing Systems* (NIPS'17)
4. Jacob Devlin,Ming-Wei Chang,Kenton Lee,Kristina Toutanova, **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*
5. A Primer on Neural Network Models for Natural Language Processing By Yoav Goldberg https://arxiv.org/abs/1510.00726

Weekly Schedule

Week	Topics Covered	Text Book Chapter
1	Introduction, Motivation of Course, Text Preprocessing, Regular Expressions Statistical Language Modeling, N Grams, Smoothing, Perplexity	2, 3
2	Sentiment Analysis using Naïve Bayes Classification, Text Classification Evaluation (Accuracy, Precision, Recall, F Measure, Multi class, Micro and Macro Average)	4
3	Vector Semantics, TF-IDF, Cosine Similarity (Sparse Vectors), Distributed word representations, Word Embeddings (WordtoVec) (Dense Vectors), Fastext, Glove	6
4	Neural Language Modeling	7
5	Recurrent Neural Networks, Vanishing Gradient Problem, LSTM, GRU , Sequence Processing with Recurrent Networks	9
Midterm 1		
6	Encoder Decoder Model, Neural Machine Translation, Natural Language Generation (NLG)	13
7	Decoding schemes, Attention in encoder decoder model, NLG Evaluation Measures (ROUGUE, BLEU), Summarization	
8	Self-Attention and Transformers, BERT	10
9	Fine tuning and Masked Language Models, Pretraining, Prompting, Reinforcement Learning from Human feedback	11
10	Sequence Labeling for Parts of Speech and Named Entities	8

Midterm 2		
11	Automatic Speech Recognition	16
12	Advanced Topics	
13		
14		

Tentative Grading Criteria

1. 5 Assignments (15%)
2. 3 best out of 4 Quizzes (12%)
3. Midterm Exams (23%)
4. Final Exam (50%)

Course Policies

1. Quizzes will be announced.
2. No makeup for missed quiz or assignment.
3. 80% attendance

Plagiarism in Assignments

In writing up your assignments and in answering questions in exams, be as clear, precise, and concise as you can. **Understandability will be an important factor in grading**.

Academic Integrity: All work MUST be done individually. Any copying of work from other person(s) or source(s) (e.g. the Internet) will automatically result in at least an F grade in the course. It does not matter whether the copying is done in an assignment, quiz, midterm exam, or final exam, it will be considered equally significant.