UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie

School of Information Technology
Department of Computer Science

Natural language processing (COS 760)

Lecturers:
Dr Abiodun Modupe; BDS Coordinator, CS. UP.
Prof Vukosi Marivate; Chair of Data Science, CS. UP.

Last Revision: February 14, 2025

Contents

1 Introduction **1**
1.1 Welcome 1

2 What is this course about? **1**

3 Reference Material **2**

4 Educational approach **2**

5 Assessment **3**
5.1 Assessment plan 3
5.2 Assessment criteria 3
5.2.1 Assignments [50%] 3
5.2.2 Final Project [50%] 3
5.2.3 Assessment Policy 4

6 Course Objective **4**

7 Plagiarism **4**

8 UP guidance on Generative AI **4**

9 Course Outcome **4**
9.1 Schedule 5
9.2 Additional Material 6

1 Introduction

1.1 Welcome

Welcome to COS761. Natural language processing involves making human language understandable for computers. In today's world, natural language processing has become an essential part of our daily lives, changing how we create and understand information. Natural language processing (NLP) is a crucial part of artificial intelligence (AI), which models how people share information. In recent years, machine learning and deep learning methods have achieved very high performance in many NLP tasks. In this course, students are introduced to advanced neural networks for NLP, including automatic machine translation widely used on the Web and on social networks. Text classification helps prevent our email inboxes from being flooded with spam. Search engines have moved beyond string matching and network analysis to a high degree of linguistic sophistication. Dialogue systems provide an increasingly common and effective way to obtain and share information.

These diverse applications are based on a common set of ideas, drawing on algorithms, linguistics, logic, and statistics. The goal of this text is to provide a survey of the fundamental concepts and principles of natural language processing. The technical exploration begins in the next chapter. The chapter identifies some high-level themes in contemporary natural language processing and advises the reader on how best to approach the topic.

(a) Dr. Abiodun Modupe

(b) Prof. Vukosi Marivate

2 What is this course about?

Natural language processing (NLP), or computational linguistics, is one of the most important technologies of the information age. NLP applications are ubiquitous, as they are used in various areas such as web search, advertising, emails, customer service, language translation, virtual agents, medical reports, and politics. In the 2010s, deep learning (or neural network) approaches achieved high performance in various NLP tasks by using single neural models that did not need traditional, task-specific feature engineering. Scaling up large language models like ChatGPT led to significant advances in the 2020s. This course offers

a comprehensive introduction to the fundamentals of machine learning and deep learning for NLP and the latest research on large language models (LLMs). By attending lectures, completing assignments, and a final project, students will acquire the essential skills to create, implement, and build neural network models using Python with the PyTorch, Keras, or TensorFlow framework.

Learning presumed to be in place

- Python proficiency is a crucial skill for any data scientist or machine learning engineer. All class assignments will be in Python. If you have a lot of programming experience but in a different language (e.g., C/C++/Matlab/Java/R/Javascript), you should still manage to follow, but Python is highly recommended because of its extensive libraries and ecosystem, which we shall rely on throughout the duration of the course.
- Calculus, linear algebra You should be comfortable taking (multivariable) derivatives and understanding matrix/vector notation and operations.
- Basic Probability and Statistics
  You are expected to know the basics of probabilities, Gaussian distributions, means, standard deviations, etc.
- Foundations of Machine Learning We will be formulating cost functions, taking derivatives, and performing optimisation with gradient descent. If you already have basic machine learning and/or deep learning knowledge, the course will be easier.

This module is closely related to the AI modules.

3 Reference Material

The following texts are useful, but none are required. All of them can be read free online.

Title:	Speech and Language Processing (Third Edition draft: 2024 pre-release)
Author:	Dan Jurafsky and James H. Martin.
Download:	Website

Title:	Natural Language Understanding with Distributed Representation
Author:	Kyunghyun Cho 2015.
Download:	Website

4 Educational approach

The lecturer will provide material for the class through lecture slides. As part of teaching and learning, students will receive information before classes to read through to familiarise themselves with some of the content.

Students are required to attend lectures, practical sessions, and tutorial sessions. Mentor sessions are compulsory for students who are invited and are available to students who

would like guidance on a specific aspect of the module content.

The module follows a continuous assessment model and therefore it is imperative that the student stay up-to-date. Students who apply a continuous work schedule are more likely to succeed than students who work sporadically and tend to work in bursts before crucial deadlines and examination opportunities. Therefore, it is recommended to continue to take advantage of the available opportunities to ensure that you understand the work. It is hard work to stay up to date, but much harder to catch up.

5 Assessment

COS761 is a continuous assessment module and the continuous evaluation process of the modules are described in the sections that follow.

5.1 Assessment plan

Students will receive four assessment opportunities based on the topics covered in the modules during the semester.

5.2 Assessment criteria

5.2.1 Assignments [50%]

There will be four assignments throughout the semester. These assignments are designed to enhance both your theoretical understanding and your practical skills. All assignments consist of written questions and programming tasks that provide a mix of theoretical and practical challenges.

- **Credit**
    1. Assignment 1 (10%)
    2. Assignment 2 (10%)
    3. Assignment 3 (15%)
    4. Assignment 4 (15%)
- **Deadlines** All assignment due dates will be posted on Click-up on the day of release, and students are expected to submit via the Turnitin submission provided for each of the assignments (i.e., before 23:59). Students are encouraged to submit their assignment before the deadline. **Do not email us your assignments.**

5.2.2 Final Project [50%]

The list of project topics will be available on Blackboard one day after the first class, and students will be able to apply in groups of two or more based on the in-depth skills learnt in class. This will allow students to choose a project that aligns with their interests and strengths, as well as collaborate with classmates to produce high-quality work. Students are encouraged to think about potential topics early in the semester to ensure that they have enough time to develop a strong proposal, milestone, poster, and final report.

- **Credit** Credit for the final project is divided into the following categories:

1. Project proposal (10%)
2. Project presentation (10%)
3. Project Poster (10%)
4. Project Report (20%)

- **Deadlines** TOA

5.2.3 Assessment Policy

The student must obtain a final mark of at least 50% to pass the module.

6 Course Objective

1. The primary goal of this module, COS760, is to help students understand how NLP is used in a variety of real-world domains such as information retrieval, sentiment analysis, language translation, and document summarization.
2. Help students learn different techniques and algorithms used in NLP, as well as gain hands-on experience in the implementation of NLP applications. The module provides a comprehensive overview of current trends and challenges in the field of natural language processing.

7 Plagiarism

Plagiarism is a serious form of academic misconduct. It involves both appropriating someone else's work and passing it off as one's own work afterwards. Thus, you commit plagiarism when you present someone else's written or creative work (words, images, ideas, opinions, discoveries, artwork, music, recordings, computer-generated work, etc.) as your own. Only hand in your own original work. Indicate precisely and accurately when you have used the information provided by someone else. Referencing must be done in accordance with a recognised system. Indicate whether you have downloaded information from the Internet. For more details, visit the following websites: `https://www.up.ac.za/news/post_1956240-important-information-about-rules-regarding-plagiarism`.

8 UP guidance on Generative AI

Generative AI for Teaching and Learning Guidelines

1. Lecturer's Guide (2025)
2. Student's Guide (Complete)

9 Course Outcome

The learning outcomes are broken down as follows:

1. Students learn how to implement basic computer programming techniques on large amounts of text, as well as how to automatically extract essential words and phrases that characterise a work's style and content. These abilities are required for activities such as data analysis, natural language processing, and information retrieval. By understanding these strategies, students will acquire valuable insight from text data to improve the efficiency of numerous procedures.

2. Students become acquainted with machine learning algorithms (e.g., machine learning algorithms (e.g., Naive Bayes, SVM, Decision Trees) and deep learning models (e.g., RNNs, LSTMs, GRUs, transformers, and attention mechanisms) as well as understand how these models are applied to tasks such as machine translation, language generation, and named entity recognition.

3. Students learn how to handle raw text data, such as cleaning, tokenization, stopword removal, and vectorisation (e.g., TF-IDF, Word2Vec, GloVe). These skills are essential for natural language processing tasks such as sentiment analysis, text classification, and language modelling. Understanding how to pre-process raw text data is crucial for extracting meaningful insights and patterns from unstructured text.

4. Students learn about ethical considerations and bias in NLP, e.g. Facilitate understanding of the ethical implications of NLP technologies, including bias in language models, fairness, privacy concerns, and the social impact of NLP tools. This knowledge helps students develop critical thinking skills and make informed decisions when working with NLP technologies in various fields such as healthcare, finance, and social media. Understanding these ethical considerations is crucial to creating responsible and inclusive NLP applications that benefit society as a whole.

9.1 Schedule

Lecture notes will be uploaded one day before or after each class. The lecture notes will cover approximately the first half of the course content, along with supplementary materials beyond the lectures. Students are encouraged to review the lecture notes before class to improve their understanding of the material and come prepared with any questions they may have. In addition, the lecture notes will serve as a valuable study resource for assignments and the final project throughout the semester.

Date	Type	Topic	Resources
12/02	Class 1	Intro. to NLP	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 1)
19/02	Class 2	Regular Expressions, Tokenization, Edit Distance (Text Processing)	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009) (Ch. 2)
26/02	Class 3	N-gram Language Models	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 3)
05/03	Class 4	Vector Semantics & Embeddings	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 6)
12/03	Class 5	Naive Bayes & Logistic Regression (Text classification & Sentiment)	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 4–5)
19/03	Class 5	Neural Networks, RNNs & LSTMs	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 7–8)
26/03	Class 6	Project Instruction & Transformers	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 9)
02/04	Class 7	Large Language Models	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 10)
09/04	Class 8	Masked Language Models	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 11)
23/04	Class 9	Model Alignment, Prompting & Context learning	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch.12)
30/04	Class 10	NLP Applications I (Machine Translation)	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 13)
02/05	Class 11	NLP Applications II (Q & A, Information Retrieval, and RAG)	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 14)
07/05	Class 12	NLP Applications III (Dialogue Systems)	Natural Language Processing with Python by Bird, S., Klein, E., & Loper, E. (2009). (Ch. 15)
14/05	Class 13	Ethical consideration in NLP systems I	Ethical by Design: Ethics Best Practices for Natural Language Processing
21/05	Class 14	Ethical consideration in NLP systems II	
28/05	Class 15	Project Presentation	Venue TOA

9.2 Additional Material

Additional material and references for students on each of the above topics is listed below.

- Introduction to NLP includes the following
    1. What is natural language processing?
    2. What are the features of natural language?
    3. What do we want to do with NLP?
    4. What makes it difficult?
    5. Building a rule-based classifier
    6. Training a bag-of-words classifier
    7. Reading Material
        - Examining Power and Agency in Film (Sap et al. 2017)
- Regular Expressions, Tokenization, & Edit Distance
    1. Regular expressions
    2. Words, Corpora, Word/Subword Tokenization
    3. Word Normalization,Lemmatization & Stemming
    4. Sentence Segmentation & Edit Distance
    5. Reading Material
        - Tokenization Falling Short: On Subword Robustness in Large Language Models (Chai, Yekun, et al. 2024)
        - Unigram Models for Subword Segmentation (Kudo 2018)
        - Between words and characters: A Brief History of Open-Vocabulary Modelling and Tokenization in NLP. (Mielke et al. (2021)
- N-gram Language Models & Embedding
    1. N-Grams
    2. Training, Testing & Evaluating language models
    3. Language Model Perplexity
    4. Cosine Similarity Measure
    5. TF-IDF: Weighing terms in the vector
    6. Reading Material
        - Ts-grams: Defining generalized n-grams for information retrieval (Järvelin, Anni et al. 2007)
        - Text mining using n-grams (Schonlau, M. and Guenther, N., 2017.)
        - Applications of n-grams in textual information systems (Robertson, A. M., & Willett, P. (1998)
        - What makes my model perplexed? a linguistic investigation on the perplexity of neural language model perplexity (Miaschi et al. (2021)
        - Learning similarity with cosine similarity ensemble (Xia, P., Zhang, L., & Li, F. (2015)
        - A novel TF-IDF weighting scheme for effective ranking
- Ethical Considerations in NLP Systems
    1. Motivation examples and practical tools for assessing adversarial NLP systems

2. Significant background on NLP (e.g., ML & DL) algorithmic bias with high-level overview
3. Reading Material
   - Ethics Best Practices for Natural Language Processing (Leidner, J. L., & Plachouras, V. (2017, April).))
   - Beyond fair pay: Ethical implications of NLP crowdsourcing (Shmueli, Boaz, et al 2021)
   - A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope by Partha Pratim Ray (2023)
   - Social & Ethical Considerations in NLP Systems by Yulia Tsvetkov from Carnegie Mellon University (CMU)
4. Bias in NLP
   - What are the biases in my word embedding? AIES (2019) by Swinger, De-Arteaga, et al.
   - De-Arteaga et al. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. FAT (2019)
   - Gonen, et al. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. NAACL (2019).
   - Manzini et al. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. NAACL (2019).