UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

**School of Information Technology**

**Department of Computer Science**

COS801-2022
**Project Description**

Lecturer: Abiodun Modupe

Last Revision: 07 July 2022 (Version 1.1)

**Submission Deadline: October 29, 2022**

**Project Expectations**

One of the objectives of the COS801 modules is to get students comfortable with designing and implementing deep learning networks from various state-of-the-art (SOTA) published papers that have been used to solve real-world problems. For 2022, we are looking at published research papers in various domains to see if we can replicate the concept/methods/idea and get a better result or something similar with the same datasets. The project will be done in a group of two or a maximum of three students. Students, we will be groups and allocated to a research subject on or before August 12th, 2022. The project is the main output of this class and will represent a 40% mark for the semester test.

Each topic addressed in the project description has datasets available. You are urged to use this to get your ideas for the project and arrange a meeting to discuss them with me (including your groupmate) before the project submission deadline of October 29, 2021.

You need to emphasize on the following in your project report:

- A scalable implementation of the model/approaches/methods
- Experimental evaluations of the model/approaches/methods from the datasets and the results as compared to the SOTA methods
- You need to report your finding, limitation, and future work from your own perspective.

**Rubrics:** Available on ClickUP.

**Expected Solutions and mark allocation.**

The following breakdown are expected:

- Introduction and problem description -- 10%
- Approaches/Model/Methods/Algorithms description -- 10%
- Experimental description of the dataset and results – 5%
- Referencing and citation – 5%
- Conclusion/Discussion/Future work --5%
- Presentation: Organization and flow – 5%

**Total: 40%**

# Machine learning and Deep learning project

### 1. Automated Question Tagging on stack Overflow

- Make a multi-label classification system that automatically assigns tags for questions posted on a forum such as Stack Overflow or Quora.
- Dataset: shorturl.at/bfkt5

### 2. **Keyword/Concept identification**

- Identify keywords from millions of questions from text questions
- Dataset: shorturl.at/bfkt5

### 3. Post Authorship identification of short texts

- To use deep learning model to understand the writing style of an individual based on the given online text snippets.

- Multi-label classification of printed online text snippets or documents of various authors
- Dataset: shorturl.at/bfkt5

### 4. Explainable prediction to credit score using deep regularized neural network model

- The idea is to use convert tabular datasets into images and apply 2D CNN with max-pooling stack upon a sequence model, e.g., LSTM to find a feature representation that can be used to predict credit score.

- Dataset: 1. https://archive.ics.uci.edu/ml/index.php

  2. https://www.kaggle.com/datasets/ajay1735/hmeq-data

### 5. Named Entity Recognition (NER) in a low resource environment

- Investigate different methods that are used to improve NER for low-resourced languages such as African languages.

- Different architectures and what could be improved in the current state-of-the-art research experiments

- Dataset: https://github.com/masakhane-io/masakhane-ner/tree/main/data

**6. Data augmentation approaches for legal document analytics.**

- Data augmentation is applying transformations to given training datasets to expand them syntactically. The aim is to use domain-agnostic data augmentation techniques that operate in the input feature spaces. In the context of contract element extraction by Chalkidis et al. (2017a, b), the purpose is to explore the effectiveness of three domain-agnostic data augmentation methods, e.g., adding Gaussian noise, applying interpolation, and extrapolation of word embedding models. The performance of the augmentation methods will be tested with a sequence model, e.g., the LSTM/BLSTM contract element extraction method by Chalkidis et al., 2017b on a subset of six out of eleven contract element types considered by Chalkidis et al. that focus primarily on element types with fewer data samples in the (non-augmented) training set of Chalkidis et al.

- Dataset: Extracting Contract Elements by Chalkidis et al (2017).

**Submission Instructions and Deadline**

On or before the deadline, please submit your report, code (which needs to be running/working), and declaration of originality in pdf file format on the course ClickUp. Submissions received after the deadlines will not be considered.

**Plagiarism**

This department considers plagiarism to be a serious offence. Disciplinary action will be taken against students who commit plagiarism. For more information on plagiarism, please refer to http://www.library.up.ac.za/plagiarism/index.htm.

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.

Please ensure you write your project report in your own words. You will be required to submit, as part of this project report a declaration of originality.