INM434/IN3045: Natural Language Processing

Final coursework
Submission deadline: **May 8, 2024 17:00 hrs**

Introduction

This coursework offers the opportunity to either create or utilize NLP systems. The primary objectives of the coursework is to apply algorithms on a dataset and report on the empirical results and provide a sufficiently deeper analysis. Alternatively, the coursework can also focus on construction of a dataset and with a thorough analysis and baselines. There are a number of different areas of focus that can be considered, this includes implementation and development of algorithms, defining a new task or applying a linguistic formalism, or exploring a dataset or task in more depth. This coursework will deepen your understanding of the challenges and opportunities in NLP, whether you choose to focus on the technical side of building NLP systems or the linguistic and theoretical foundations of the field.

The coursework will comprise the design and implementation of the NLP pipeline on a standard NLP task over a real-world dataset with critical analysis and evaluation of the performance of the implementation. The students will provide a written report as a short research paper of approximately **2,000 words (for UG students who are taking IN3045) or 4,000 words (for PG students who are taking INM434)** accompanied by a **5-min long video presentation**. The students will be awarded marks for the following:

1. Identifying the research question(s) for a task that can be addressed with an NLP pipeline.
2. Designing and implementing the working NLP pipeline.
3. Demonstrating that good practice has been considered for assembling the NLP pipeline.
4. Critically evaluating the proposal.
5. Reflecting on the complete process.

The subject of the above learning outcomes follows the lectures and associated materials and exercises.

Marking information

This individual coursework, comprises 85% of your final module marks. Collaborative work is not permitted for this coursework.

Submission

Only submissions through `Moodle` will be accepted.

You are required to submit:

1. **All `python` code (as a single `.zip` package) required to run your two best-trained models on a test set**. Include in the `.zip` package, a `readme.txt` file with all the instructions on how your models should be run, including library or directory dependencies and required software versions. If possible, include the test set in the `.zip` file; otherwise, add a web-link to the `readme.txt` file stating where the test set can be downloaded from.
2. Additionally concatenate all the code into a single text file and submit this to `Moodle`.
3. A pdf report - the template will be made available through `Moodle`.
4. Supplementary material (as a `pdf` file) of up to a maximum of four A4 pages, including any relevant intermediate results and implementation details which are not explicit in the paper.

No late submissions will be accepted. You are strongly advised to submit draft versions of the coursework well before the deadline. Please don't leave the submission to the last minute!

Teamwork

Teamwork is not allowed. This coursework is exclusively individual work.

UG/PG requirements

Please note that the requirements for UG and PG students are slightly different. UG students will submit a report of 2,000 words (approx. 4 pages) with a short 5-min video presentation. PG students will submit a report of 4,000 words (approx. 8 pages) with a short 5-min video presentation.

Marking scheme

The coursework (`python` code, paper and supplementary material (or the appendix) will be marked according to the following criteria:

	Content	Distribution
Code	Syntactic correctness	5%
	Organization and clarity of comments	5%
	Appropriate use and sophistication of methods	15%
Report	Description and motivation of the problem	5%
	Background and literature review	5%
	Initial analysis of the data set including basic statistics	5%
	Summary of the NLP pipeline/algorithms with their pros and cons	5%
	Hypothesis statement	5%
	Description of the choice of training and evaluation methodology	5%
	Choice of parameters and experimental results	10%
	Analysis and critical evaluation of results	15%
	References, lessons learned and future work	5%
Video	A 5-minute presentation summarising the project	15%
	Total	100%

You will not be marked on how good the results are. What matters is that you follow a sound methodology and present clearly the problem, your method and the results, with a critical and fair comparative evaluation.

General Guidelines

This coursework will provide you with hands-on experience in conducting practical applications of NLP models. The goal of the report is to effectively communicate your findings and connect them to existing literature. You will be required to write code, run experiments on selected data, read and analyze relevant papers, and present your results through a paper that includes appropriate figures and tables. It is important to properly cite any published material used, whether found on the web or in a textbook. Always strive to explain your results in your own words.

Extenuating Circumstances

If you are not able to submit your coursework for medical reasons or any serious personal reasons beyond your control you should contact the programmes office (UG: `sst-ug@city.ac.uk`; PG: `sst-pgoffice@city.ac.uk`) and fill an extenuating circumstances form. Please note that medical certificates will be required.

Plagiarism

All submissions — both code and text have to be produced independently. Consider the following illustrative scenarios:

Acceptable: A and B discuss options for efficiently storing large vectors as required for an NLP problem.
Unacceptable: A and B collaborate on coding the solution to store feature counts.
Acceptable: B seeks clarification from A implementing the Byte Pair Encoding algorithm, and A provides a conceptual explanation.
Unacceptable: A and B divide the coursework into parts and each write their own code, then combine their solutions to complete the task.
Acceptable: A inquires about B's experience with a failure condition during coding and B describes his conceptual approach to resolve it.
Unacceptable: A or B use a pre-existing solution from a machine, an open repository or a similar assignment from another class as a starting point for their own solution.
Unacceptable: A uses the help of generative AI models for some part of the work - immediately copies and pastes the outputs into the coursework.

Please follow guidelines as directed in INM373 - Research Methods and Professional Issues. Please consult the following hyperlink for additional guidance:
https://studenthub.city.ac.uk/help-and-support/academic-integrity-and-misconduct.