University of South Wales
Prifysgol De Cymru

Faculty of Computing Engineering and Science Assessment Brief

Module Title: Deep Learning
Module Code: CS4S772
Module Leader/Tutor: Dr Mabrouka Abuhmida
Assessment Type: Practical Assessment
Assessment Title:
Weighting: 50%
Word count/duration/equivalent: NA
Submission Date: 013-01-25
Return Date: in 20 working days from the submission date

Assessment Description

Design and implement a complete Natural Language Processing (NLP) pipeline for advanced sequence-to-sequence tasks, including text summarisation, semantic search, and thematic analysis using the Sherlock Holmes dataset. The focus is on understanding the process, implementing modular steps, and critically evaluating outcomes.

Objective To write **a comprehensive report** detailing the development, findings, and results of your (NLP) pipeline.
focusing on:

- How design choices influenced performance.
- Challenges encountered at each stage.
- Insights gained from the dataset and NLP methods used.
- Suggest improvements for each component of the pipeline.

Data: Use the provided dataset on Blackboard. Sherlock Holmes Stories.zip

This collection features all the stories and novels of Sherlock Holmes by Arthur Conan Doyle. Within the Sherlock folder, you'll find multiple .txt files, each containing a unique story.

Assignment Tasks

Task 1:
Clean the Sherlock Holmes dataset to handle common text preprocessing challenges, provide a short report detailing preprocessing challenges and how they were addressed.

Task 2:

- Implement a Seq2Seq model using an Encoder-Decoder LSTM architecture for summarizing entire Sherlock Holmes stories into short summaries.
- Train the model to produce summaries of around 50-100 words.
- Evaluate the model on both using suitable score metrics based on your research.

Task 3:

- Create vector embeddings for individual paragraphs from the Sherlock Holmes dataset using pre-trained models like all-MiniLM-L6-v2.
- Store these embeddings in ChromaDB.
- Implement a semantic search system where users can query sentences or phrases, and the system retrieves semantically similar paragraphs.

Demonstrate retrieval for specific queries like:

- "Find paragraphs related to detective investigations."
- "Retrieve sections discussing Dr Watson."

Task 4:

Perform Topic Modelling on the cleaned dataset using Latent Dirichlet Allocation (LDA) and provide a brief analysis of identified topics (e.g., crime, investigation, relationships). Identify key topics across different Sherlock Holmes stories and analyse how they vary. Map the distribution of topics in individual stories and the entire corpus.

Report Structure

- Introduction: Overview of the problem, its significance, and objectives of the project.
- Methodology: Detailed description of the model architecture, data preprocessing, and training procedures.
- Results: Detailed presentation of the model's performance with appropriate statistical analysis, graphs, and charts.
- Deployment: Explanation of deploying the model using Anvil, including any challenges and solutions.
- Discussion: Critical analysis of the model's performance, comparison with existing solutions, and discussion of any limitations.
- Conclusion: Summarize key findings, lessons learned, and potential future work.
- References: Properly cited sources and literature if used.

Guidance on Format of Assessment

Note: Students are reminded **not** to include this assignment brief with the assignment submission

- Your report should be no longer 2000-2500 words.
- Include necessary graphs, tables, or visuals to support your findings.
- Clearly label all sections and subsections.
- Cite any sources or references used.

Report Submission Guidelines:

- Implement your model.
- Include the code for the main techniques (only) used in your experiments as a part of the submission.
- Report the results of your experiments in the report.
- Conduct an evaluation of the results, discussing the impact and any potential limitations.
- Conclude with a summary of your findings and recommendations for further work or potential real-world applications.

Code Submission Guidelines:

- Your code should be well-organized and commented for clarity.
- Use appropriate naming conventions for functions, variables, and classes.
- Code must be submitted as a Jupyter notebook
- Ensure the code can be executed to reproduce the results mentioned in the report.

Learning Outcomes Assessed

- To demonstrate knowledge and comprehension in the explanation of deep learning methods, tools and models.

- To demonstrate discernment in the effective application of deep learning to design and implement a problem dependent solution utilising appropriate deep learning methods, tools and models.

Marking Criteria/Rubric

Note: All grades are provisional until they are ratified by the exam board

Section	Criteria	Excellent (80-100%)	Good (60-79%)	Fair (40-59%)	Poor (0-39%)
1. Data Preprocessing & Report (20%)	Data Cleaning	Thorough cleaning addressing a wide range of challenges. Clearly justifies choices.	Adequate cleaning with most common steps implemented. Justification provided.	Basic cleaning with some important steps missing or not well justified.	Limited cleaning with significant steps missing. Little/no justification.
	Report - Preprocessing	Detailed and insightful report on challenges and solutions. Demonstrates strong understanding.	Good description of challenges and solutions. Shows understanding of importance.	Adequate description with some explanation of challenges and solutions.	Limited or superficial discussion of preprocessing.
2. Text Summarization (25%)	Seq2Seq Model Implementation	Correctly implements Encoder-Decoder LSTM. Code is well-structured and documented.	Implements the core model with minor errors. Code is reasonably structured.	Attempts to implement the model, but with significant errors.	Fails to implement a functional model.
	Model Training & Evaluation	Effectively trains the model, achieving good performance (BLEU score). Evidence of hyperparameter tuning.	Trains the model with reasonable results. Some evidence of tuning.	Trains the model with limited success. Limited tuning.	Fails to train/evaluate effectively.
	Report - Summarization	Comprehensive discussion of model, training, and results. Insightful analysis of performance and limitations.	Good discussion of model and training. Provides some analysis of results.	Adequate description of model and training. Limited analysis.	Poor description of model/training. Little/no analysis.
	Advanced Challenge (Extra Credit - 5%)	Successfully implements an advanced technique (e.g., attention) with	Attempts to implement an advanced technique with partial success.	Attempts an advanced technique with significant errors.	Does not attempt the challenge.

		clear improvement.			
3. Semantic Search (25%)	Embedding Generation & Storage	Correctly generates embeddings using all-MiniLM-L6-v2 and stores them in ChromaDB.	Generates embeddings and stores them with minor errors.	Attempts to generate embeddings and use ChromaDB, but with errors.	Fails to generate embeddings or use ChromaDB effectively.
	Semantic Search Implementation	Implements a functional system that accurately retrieves semantically similar paragraphs.	Implements a basic system with limitations in accuracy.	Attempts to implement search, but with significant errors.	Fails to implement a functional system.
	Query Demonstration	Clearly demonstrates functionality with specific queries and provides insightful analysis of results.	Demonstrates functionality with some analysis of results.	Limited demonstration with little analysis.	Fails to demonstrate functionality effectively.
	Report - Semantic Search	Thorough discussion of embedding generation, ChromaDB, and search implementation. Insightful analysis of performance.	Good discussion of embedding generation and search. Some analysis of results.	Adequate description of embedding generation and search. Limited analysis.	Poor description of embedding generation and search. Little/no analysis.
4. Topic Modeling (15%)	LDA Implementation	Correctly implements LDA for topic modeling.	Implements LDA with minor errors.	Attempts to implement LDA with significant errors.	Fails to implement LDA effectively.
	Topic Analysis	Clear and insightful analysis of identified topics, including interpretation and relevance.	Good analysis of topics with some interpretation and discussion of relevance.	Basic analysis of topics with limited interpretation.	Fails to provide meaningful analysis of topics.
	Topic Distribution & Mapping	Effectively analyzes and visualizes topic distribution across stories	Analyzes and visualizes distribution with some discussion of variations.	Basic analysis of distribution with limited visualization.	Fails to analyze or visualize distribution effectively.

		and corpus. Meaningful conclusions.			
	Report - Topic Modeling	Comprehensive discussion of LDA, topic analysis, and distribution mapping. Insightful observations.	Good discussion of LDA and topic analysis. Some interpretation of distribution.	Adequate description of LDA and analysis. Limited discussion of distribution.	Poor description of LDA and analysis. Little/no discussion of distribution.
5. Report Overall (15%)	Structure and Clarity	Exceptionally well-structured, clear, and concise. Uses appropriate language and visualizations.	Well-structured and generally clear. Uses visualizations appropriately.	Adequately structured but may lack clarity. Visualizations may be limited.	Poorly structured and difficult to follow. Limited visualizations.
	Critical Evaluation & Insights	Deep understanding of NLP methods and insightful critical evaluation of performance. Meaningful conclusions and improvements.	Good critical evaluation and suggests some improvements. Good understanding of methods.	Some critical evaluation but may lack depth. Basic understanding of methods.	Limited critical evaluation or insights. Limited understanding of methods.
	Code Quality (Included in code weighting - 40% overall)	Code is well-structured, efficient, and thoroughly documented. Good coding practices.	Code is reasonably well-structured and documented.	Code is functional but may be poorly structured or lack documentation.	Code is poorly structured, inefficient, and difficult to understand.

What happens next?

Your marked assessment should be available 20 working days after submission. However, please be advised that this may be subject to change in the event of Bank Holidays, University Closure or staff sickness. If there is something about the feedback you have been given that you are unclear about, please see your module tutor. Mabrouka.abuhmida@soutwales.ac.uk

Feedback Method

Personalised feedback will be provided via Blackboard feedback

Late Submission

Late submissions that is 5 working days late will be capped to 40%

Retrieval in the Event of Failure

You get to attempt IYR to submit an improvised vision of your submission up to ten working days after your feedback. You need to apply for this, and you only allowed to use this twice in the same academic year.

Extenuating Circumstances

https://advice.southwales.ac.uk/a2z/extenuating-circumstances

Referencing, Plagiarism and Good Academic Practice

https://advice.southwales.ac.uk/a2z/referencing-plagiarism-and-good-academic-practice

Learning Support Resources

https://studyskills.southwales.ac.uk

Your Assessment Queries

I am holding an assessment surgery session. And also, will be dealing with inquiries via Email. If I get different inquiries regarding similar issues, I will post a summary on the Module chat group.