**XJTLU Entrepreneur College (Taicang) Cover Sheet**

|  |  |  |
| --- | --- | --- |
| Module code and Title | **DTS305TC Natural Language Processing** | |
| School Title | **School of Artificial Intelligence and Advanced Computing** | |
| Assignment Title | **Coursework 2 (CW 2)** | |
| Submission Deadline | **5 pm China time (UTC+8 Beijing) on** Sat. 2nd. Nov. 2024 | |
| Final Word Count | **8000** | |
| If you agree to let the university use your work anonymously for teaching and learning purposes, please type **“yes”** here. | |  |

I certify that I have read and understood the University’s Policy for dealing with Plagiarism, Collusion and the Fabrication of Data (available on Learning Mall Online). With reference to this policy I certify that:

* My work does not contain any instances of plagiarism and/or collusion.  
  My work does not contain any fabricated data.

**By uploading my assignment onto Learning Mall Online, I formally declare that all of the above information is true to the best of my knowledge and belief.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Scoring – For Tutor Use** | | | | | | |
| **Student ID** | | | |  | | |
|  | | | | | | |
| **Stage of Marking** | | **Marker**  **Code** | **Learning Outcomes Achieved （F/P/M/D）**  **(please modify as appropriate)** | | | **Final**  **Score** |
| **A** | **B** | **C** |
| 1st Marker – red pen | |  |  |  |  |  |
| Moderation  – green pen | | **IM**  **Initials** | The original mark has been accepted by the moderator (please circle as appropriate): | | | Y / N |
|  | Data entry and score calculation have been checked by another tutor (please circle): | | | Y |
| 2nd Marker if needed – green pen | |  |  |  |  |  |
| **For Academic Office Use** | | | **Possible Academic Infringement (please tick as appropriate)** | | | |
| **Date**  **Received** | **Days late** | **Late Penalty** | **Category A** | | Total Academic Infringement Penalty (A,B, C, D, E, Please modify where necessary) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |
|  |  |  | **Category B** | |
| **Category C** | |
| **Category D** | |
| **Category E** | |

**Students**

**(Please modify where necessary)**

The assignment must be typed in an MS Word and converted to a PDF document. The document must be submitted via Learning Mall Online to the correct drop box. Only electronic submission is accepted and no hard copy submission.

All students must download their file and check that it is viewable after submission. Documents may become corrupted during the uploading process (e.g. due to slow internet connections). However, students themselves are responsible for submitting a functional and correct file for assessments.

**DTS305TC Natural Language Processing**

**Coursework 2 (Individual Assessment)**

**Due: 5:00 pm China time (UTC+8 Beijing) on** Sat. 2nd. Nov. 2024

**Weight: 60%**

**Maximum score: 100 marks (100 individual marks)**

**Assessed learning outcomes:**

**C. Implement deep learning models and evaluate them based on performance metrics.**

**D. Develop skills of using NLP models and techniques in real-world applications.**

**Overview**

Document classification is a core NLP task that involves automatically categorizing written content into a predefined set of classes or categories. This process is crucial for managing the vast amounts of textual data generated daily across various domains, including news, legal documents, medical records, and online content. The key aspects of document classification include: Text Representation, Feature Extraction, Model Selection, Deep Learning Approaches, Performance Evaluation, and so on.

This task faces challenges such as handling imbalanced datasets, dealing with the nuances of human language including sarcasm and context, and adapting to domain-specific vocabularies and terminologies.

**Tasks**

You are required to use the slides and Internet resources to learn the detailed knowledge of document classification problem, and use the python programming language to complete one document classification report.

1. **Background Knowledge (10 Marks)**
   1. Please provide 3 real-life application scenarios that require document classification methods. (**6 Marks**)
   2. Please analyze why document classification methods, rather than other nature language processing methods, are the most suitable for these 3 application scenarios? (**4 Marks**)
2. **Data Collection (15 Marks)**

Please collect three text datasets for each of the document classification scenarios given in Task 1, with a requirement of no less than 2000 samples per dataset for classification. And provide a 100 word explanation for each dataset, including but not limited to statistical analysis of data samples, feature analysis, and other related content. (**5 Marks/dataset x 3=15 Marks**)

1. **Algorithm Design (20 Marks)**
   1. Please select 2 traditional document classification algorithms and 1 deep learning algorithm for the datasets in Task 2, and provide pseudo-code for each classification algorithm. (**5 Marks/algorithm x 3=15 Marks**)
   2. Please compare the three algorithms in 3.(1) and analyze their advantages and disadvantages. (**5 Marks**)
2. **Algorithm Implementation (30 Marks)**

Please use Python to implement the system development for each classification algorithm designed in Task 3. System development includes but is not limited to data reading (2 marks), feature processing (3 marks), classification algorithms (3 marks), results printing (2 marks), and other parts. Please submit the Python implementation of each algorithm as a separate \*.py file. Please paste the Python code of the classification algorithm function in the report. (10 Marks/algorithm x 3 = 30 Marks)

1. **Results Analysis (20 Marks)**
   1. Please provide the classification results of each algorithm for three different sample datasets, including accuracy (3 Marks), recall (3 Marks), and F1 score (4 Marks).
   2. Please discuss the different performances of different classification algorithms in accuracy (3 Marks), recall (3 Marks), and F1 score (4 Marks).
2. **Report Quality (5 Marks)**

This coursework not only assesses everyone's algorithm comprehension and implementation abilities, but also exercises your professional knowledge in terms of terminology, algorithm design expression, experimental result presentation, and the readability of function variables and function logic. Please be cautious about the quality of the report. (5 Marks)

**Submission**

You must submit the following files:

* A ***Student\_ID***.***pdf*** file contains a cover letter with your group member information, and all the task report content.
* A ***Student\_ID.zip*** file contains your program implementation and output files, such as ***dataset1.csv, dataset2.csv, dataset3.csv, algorithm1.py, algorithm2.py, algorithm3.py, accuracy.csv, recall.csv, F1.csv****.*

**End of Coursework**