**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 1 (CW 1)** | |
| Submission Deadline | **5 pm China time (UTC+8 Beijing) on** Sat. 19th. Oct. 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 1 (Group Assessment)**

**Due: 5:00 pm China time (UTC+8 Beijing) on** Sat. 19th. Oct. 2024

**Weight: 40%**

**Maximum score: 100 marks (80 % group marks + 20 % individual marks by peer assessment)**

**Groupings: Each group consists of 4-5 students. The detailed grouping table is published in the group assessment section.**

**Assessed learning outcomes:**

1. **Understand the basic concepts and techniques of Natural Language Processing.**
2. **Apply statistical and machine learning techniques to process and analyse large-scale textual data.**

**Overview**

Document clustering is an unsupervised learning task in NLP that aims to group a collection of documents into clusters based on their content similarity, without prior knowledge of the groupings. Unlike document classification, which assigns predefined categories to documents, clustering discovers the inherent groupings within the data. This task is essential for exploratory data analysis, organizing large document sets, and identifying patterns or themes that may not be immediately apparent. The key aspects of document clustering include: Text Representation, Feature Extraction, Clustering Algorithms, and Similarity Metrics.

**Tasks**

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

1. **Background Knowledge (10 Marks)**
   1. Please provide 3 real-life application scenarios that require document clustering methods. (**6 Marks**)
   2. Please analyze why document clustering methods, rather than other natural 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 clustering scenarios given in Task 1, with a requirement of no less than 2000 samples per dataset for clustering. 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 three different document clustering algorithms for the datasets in Task 2, and provide pseudo-code for each clustering 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 clustering algorithm designed in Task 3. System development includes but is not limited to data reading (2 marks), feature processing (3 marks), clustering 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 clustering algorithm function in the report. (**10 Marks/algorithm x 3 = 30 Marks**)

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

One of the group members must submit the following files:

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

**7. Peer Review**

Review your peers based on the project contribution. This will be done on LMO anonymously, each of the group members should log in to their LMO account and submit the marks individually. Marks should be submitted within a week after the group work submission is done. Peer review rubrics are attached in the Appendix.

**Appendix**

Table 1 Peer Review Rubrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Marks | 4 | 3 | 2 | 1 | 0 |
| Contributions (20%) | Routinely provides useful ideas when participating in the group discussion. A leader who contributes a lot of effort. | Usually provides useful ideas when participating in the group discussion. A strong group member who tries hard! | Sometimes provides useful ideas when participating in the group discussion. A satisfactory group member who does what is required. | Rarely provides useful ideas when participating in the group discussion. May refuse to participate. | No contribution or no submission. |
| Problem- solving (20%) | Actively looks for and suggests solutions to problems. | Refines solutions suggested by others. | Does not suggest or refine solutions, but is willing to try out solutions suggested by others. | Does not try to solve problems or help others solve problems. Lets others do the work. | No contribution or no submission. |
| Attitude (20%) | Is never publicly critical of the project or the work of others. Always has a positive attitude about the task(s). | Is rarely publicly critical of the project or the work of others. Often has a positive attitude about the task(s). | Is occasionally publicly critical of the project or the work of other members of the group. Usually has a positive attitude about the task(s). | Is often publicly critical of the project or the work of other members of the group. Is often negative about the task(s). | No contribution or no submission. |
| Focus on the task (20%) | Consistently stays focused on the task and what needs to be done. Very self-directed. | Focuses on the task and what needs to be done most of the time. Other group members can count on this person. | Focuses on the task and what needs to be done some of the time. Other group must nag, remind to keep this person on task. | Rarely focuses on the task and what needs to be done. Lets others do the work. | No contribution or no submission. |
| Working with others (20%) | Almost always listens to, shares with, and supports the efforts of others. Tries to keep people working well together. | Usually listens to, shares, with, and supports the efforts of others. Does not cause "waves" in the group. | Often listens to, shares with, and supports the efforts of others, but sometimes is not a good team member. | Rarely listens to, shares with, and supports the efforts of others. Often is not a good team player. | No contribution or no submission. |

**End of Coursework**