**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** Fri. 31. Oct. 2025 | |
| Final Word Count | **<=5000** | |
| 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.
* My work is not generated by AI.

**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** |  |
| 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 PDF document. The document must be submitted via Learning Mall Online to the correct Dropbox. 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 Fri. 31. Oct. 2025**

**Weighting: 40%**

**Maximum score: 100 marks (80 % group report + 20 % individual 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**

Text retrieval is a branch of information retrieval that refers to the technology of locating relevant documents from a text collection using keywords, semantics, and other content. Its core is natural language retrieval, which does not require manual indexing and directly matches text content such as the full text, abstract, or title. The technical implementation relies on inverted indexes, vector space models, and similarity calculations, and is mainly used in applications such as search engines and digital libraries.

**Tasks**

You are required to use the slides and Internet resources to learn the detailed knowledge of Information Retrieval System, and use the Python programming language to complete one NLP report.

1. **Background Knowledge (10 Marks)**

Write the following content in text form in the report.

* 1. Provide three real-life application scenarios that require information retrieval methods. (**6 Marks**)
  2. Analyze why information retrieval methods, rather than other natural language processing methods (document classification or document clustering), are the most suitable for these three application scenarios. (4 **Marks**)

1. **Algorithm Design (20 Marks)**

Write the following content in text form in the report.

* 1. Provide two basic processes for an information retrieval system. (**5 Marks/system x 2=10 Marks**)
  2. Provide pseudocode for the algorithms used in the information retrieval systems in 2(1). (**5 Marks/algorithm x 2=10 Marks**)

1. **System Implementation (40 Marks)**

Use Python to implement the system described in Section 2 with the following functions, and include the Python code in your report.

1. Main function: control the startup and flow of the entire information retrieval system. (**5 Marks**)
2. User input function: allow users to continuously input text to be retrieved in the console window. (**5 Marks**)
3. Database input function: read a local text library from a document folder. (**5 Marks**)
4. Text preprocessing function: preprocess the documents read in step (3). (**5 Marks**)
5. Information retrieval algorithm 1: use algorithm 1 to retrieve user input and provide the top five candidate results. (**5 Marks**)
6. Information retrieval algorithm 2: use algorithm 2 to retrieve user input and provide the top five candidate results. (**5 Marks**)
7. Output function: output the results of retrieval algorithm 1 and retrieval algorithm 2. (**5 Marks**)

(8) Fusion output function: use the weighted merge method to fuse the search results of algorithm 1 and algorithm 2, and print the new ranking results. (**5 Marks**)

1. **Results Analysis (20 Marks)**

Test your system and record the results; write the following content in text form in the report.

1. Test the developed information retrieval system using five different query examples. (**5 Marks**)
2. Use recall to analyze the three different retrieval results (algorithm 1, algorithm 2, and weighted fusion). (**5 Marks**)
3. Use precision to analyze the three different retrieval results (algorithm 1, algorithm 2, and weighted fusion). (**5 Marks**)
4. Use nDCG to analyze the three different retrieval results (algorithm 1, algorithm 2, and weighted fusion). (**5 Marks**)
5. **Conclusion (10 Marks)**

Write the following content in text form in the report.

1. Describe how your designed and implemented information retrieval system addresses the three application scenarios in Section 1. (**5 Marks**)
2. Report quality, including report format, code quality, and references. (**5 Marks**)

**Submission**

One group member must submit the following files:

* A PDF file named ***Group\_ID.pdf*** containing a cover letter with your group members’ information and all task report content.
* A ZIP file named ***Group\_ID.zip*** containing your program implementation and output files (e.g., ***dataset, IRS.py, precision.csv, recall.csv, nDCG.csv***).

**7. Peer Review**

Review your peers based on their project contributions. This will be done anonymously on LMO; each group member must log in to their LMO account and submit marks individually. Marks must be submitted within one week after the group work submission.

**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**