The Hong Kong University of Science and Technology

PG Course Syllabus

Digital Humanities – Text Mining for Humanities

HUMA 5630

Name: Dr. Steve MA

Email: hmhma@ust.hk

Office Hours: Mondays 14:30-16:30, Room 3351 (by appointment)

Interdisciplinary studies and Digital Humanities will give you a boost for job hunting and PhD applications!

Prerequisites

1. This course does not require any programming background. However, you should be at least familiar with your operating system and know how to perform common tasks such as starting up and shutting down your computer, opening files, and installing software.
2. You are required to bring your own laptop. Although the course will take place in the Computer Barn, files on the Computer Barn computers will not be saved after each restart, so they can only serve as a backup option and may not meet all the course requirements.
3. We recommend bringing a Windows laptop. MacBooks are acceptable, but Android tablets, iPads, or computers running HarmonyOS will not work.
4. Relax. While we will cover basic Python syntax, the purpose is to familiarize you with Python code. Python syntax will not be tested through exams or quizzes.
5. **We will provide a complete set of code that you can use in your future work and research. You may also use this code to complete your final group project.**

Requesting Help

1. If you encounter programming problems, you can request help from the instructor via email. To help us assist you more effectively, please include your name, the course you are taking (HUMA 5630 Digital Humanities), and a detailed description of your issue. If you encounter Python error messages, please also include both your code and the error messages in the email as well.
2. We regret to inform you that our division cannot provide any TA familiar with Digital Humanities technologies for our course. This means that it may take several working days for me to respond to your help requests. Thank you for your understanding.
3. You can also use AI tools such as ChatGPT, Grok, or Deepseek to help resolve your issues immediately. Grok is available in Hong Kong, and Deepseek is available both in Mainland China and Hong Kong. However, please note that AI assistants can make mistakes. If you are unsure about the AI's response, you can contact me to confirm whether it is correct.

4. Students with disabilities or special needs are encouraged to contact the instructor to arrange appropriate accommodation.

Assessment and Grading

The following assignments will count toward your final grade and be included in the assessment:

Mini Presentation (15%)

Each student is required to complete one mini presentation:

1. The topic of the presentation should be a database/corpus/dataset in your area of interest. Each presentation should not exceed 10 minutes. Using slides is not required.
2. Please register for your mini presentation time through the following link. The registration deadline is September 12:
   https://docs.google.com/spreadsheets/d/1bGZ623Fe-35DxQIsCc7k538ItCgNCEiiM3dM1aL1p08/edit?usp=sharing
3. If you cannot register for the mini presentation before the deadline, the instructor and TA will randomly assign you a time slot.
4. To ensure fairness, students who choose earlier presentation dates will receive bonus points, especially for the Week 2 slots.
5. Any changes to the presentation time after the register deadline must be communicated to the instructor by email.
6. You will receive your group presentation grade before the end of the semester.
7. Any application for special circumstances should be submitted by email with official proof recognized by the university. Thank you.

Grading criteria for Mini Presentation:

Developing	Accomplished	Exemplary
Does not introduce a database/corpus/dataset, or fails to explain the basic information and purpose of the database, or is seriously under/over the 10-minute limit	Introduces the basic content and purpose of the database/corpus/dataset, and provides information on how to access or download it	Introduces the database while also presenting research results using it, or provides a detailed explanation of the database's background and significance

Group Project (55%)

Students need to form groups of 2-4 people and complete the following two steps:

1. Each group needs to give one group presentation. (30%) For the group presentations, each group is required to report on their Digital Humanities project. Each presentation should not exceed 20 minutes. Please register your group members. The registration deadline is October 3.
   https://docs.google.com/spreadsheets/d/1bGZ623Fe-35DxQIsCc7k538ItCgNCEiiM3dM1aL1p08/edit?usp=sharing
2. Each group must also complete one peer review for another group. (5%) The peer review should give constructive comments to the group being reviewed. Each peer review presentation should not exceed 5 minutes. The arrangement of peer reviews will be assigned by the instructor after Week 10.

3. After the presentation, you need to submit the project-related materials within two weeks. The required materials may vary depending on the type of project. You can refer to the "Feasible Topic Examples" section in this part, or consult the instructor by email.

*Please submit your assignments **by email, not via Canvas**. My email is hmhma@ust.hk. There is no need to submit a hard copy. For presentations that require peer review, please copy your reviewing classmates when you send your slides and other materials to me.*

The group project grade is divided into the following parts: group presentation 20%, project-related materials submitted within two weeks 20%, and peer review 15%.

Grading criteria for the Group Project:

Criterion	Developing	Accomplished	Exemplary
Academic Quality 50%	Topic is unrelated to digital humanities, or violates the university's academic integrity and ethics policy, or unfairly and inappropriately uses AI tools, or chooses an inappropriate/offensive topic	Project structure is basically complete, with its own research conclusion	Project includes a complete introduction, literature review, research methods used, research process, and conclusion; the research makes contribution to humanities
Knowledge and Skills 25%	Does not use digital humanities techniques, or contains serious factual errors	Able to use digital humanities techniques to complete the research, with appropriate dataset and methods	Able to develop custom digital humanities tools or code for research according to research needs
Innovation 25%	No original viewpoint	Shows thoughts and insights on the topic	Critically engages with the topic and connects it to other academic research in the field

Feasible Topic Examples:

1. A complete digital humanities research. For example, using a fine-tuned NER model to find entities in the text, and then study and make conclusions. You need to upload the dataset used in the research and the Python scripts used in the research process to a dedicated GitHub repository, and submit the access link of the repository, and also the research essay.

2. A high-quality digital humanities dataset. You need to upload the dataset and the Python scripts used when processing the dataset to a dedicated GitHub repository, and submit the access link of the repository, and also the essay explaining the dataset and testing the dataset's performance.
3. A digital humanities GUI software. You need to upload the source code of the software to a dedicated GitHub repository, and submit the access link of the repository, and also the essay explaining the functions of the GUI software.
4. Projects in GIS, visualization, cultural heritage preservation, etc., that are not covered in class, can also be considered if suitable.

Notes:

1. If you do not register for the group presentation before the deadline, the instructor will randomly assign you to a group and give you a time slot.
2. With the unanimous agreement of group members, not everyone is required to stand on stage for the final presentation. Members can also contribute by preparing materials, creating slides, and other tasks.
3. To avoid disputes, any changes to the presentation time or group members must be communicated by email to the instructor. In particular, if you have concerns about the workload of your group members, please also contact us by email rather than telling us directly in class.
4. Submissions after the deadline will not be accepted.
5. The standard word count for the essay is 500 words. However, to facilitate potential academic conference or journal submissions, it is also acceptable to submit it following the final paper standard. Nevertheless, group members will not receive a higher grade for the extra word count.
6. According to the university's guidelines on generative AI use, you may use generative AI in this assignment, but only for the purpose of stimulating creativity. Any use of generative AI must be clearly marked or explained, including which AI tool was used and which parts of the writing were assisted by the AI tool.
   https://cei.hkust.edu.hk/en-hk/system/files?file=hkust_policy_principles_for_genai_for_tl_student_version.pdf&check_logged_in=1
7. For fairness, in accordance with university requirements, you must maintain academic integrity and comply with the university's Academic Honor Code. Violations will be handled by the Academic Registry.
   https://registry.hkust.edu.hk/resource-library/academic-honor-code-and-academic-integrity
8. If you are not sure about your final essay topic, you may schedule an office hour or contact the instructor by email.
9. Considering that there may be no time in the final week to teach how to upload code and datasets to GitHub, it is acceptable to ask the instructor for assistance with the upload, and this will not affect your final grade.

Attendance (20%)

Your participation in class is very important, and we encourage you to take part in every class.

Attendance:

1. You need to attend every class because missing sessions may make it difficult to keep up with the rest of the course.

2. Missing one Lecture may result in a 5% deduction from your Attendance grade.
3. During the Add & Drop week, we will not take attendance.

Class participation:

1. We encourage participation in class. However, we regret that, since we do not have a TA, we are unable to track students' contributions in class or award points for active participation in order to ensure the smooth running of the class.
2. As long as you do not abuse our services, any help requests during or after class will not affect your grade.

Leave of absence:

1. According to university policy, any leave request must be supported by official documents, such as a medical certificate from a hospital or proof of attending an event or academic conference. Please send these documents by email to the instructor. Showing the documents during class or office hours will not be valid.
2. If your leave affects your group presentation, it will also affect your own group presentation grade.
3. All leave requests must be submitted before you are absent. According to university policy, we cannot accept leave requests after you are already absent.

Office Hours:

1. To be fair to everyone, communication with the instructor outside of class, such as meeting in Starbucks or during booked office hours, will not affect your grade.
2. However, we still welcome your questions at any time and will do our best to support your success.

Paper Publication Assistance Service

If you would like the instructor's assistance and guidance in publishing your digital humanities work in academic conferences or journals, please let me know by email. I am more than happy to help you. However, whether the paper can be published ultimately depends on its quality.

Course Outline and Schedule

** The course schedule may be adjusted according to actual needs.*

Week 1 5 Sep

Welcome and Introduction

Week 2 12 Sep

Welcome to Python World

- By learning Python, you will master one of the most powerful research tools in the age of artificial intelligence, and gain proficiency in the core techniques of mainstream digital humanities research.

Week 3 19 Sep

Optical Character Recognition

- After learning OCR skills, you could **bring ancient manuscripts to life**, turning dusty old texts into searchable digital treasures. You could even **track handwriting quirks or hidden annotations**, uncovering little secrets from the past that were waiting to be discovered.

Week 4 26 Sep

Guset Speaker (TBA)

Week 5 3 Oct

Word Segmentation

- After learning word segmentation, you can **unlock the structure of texts**, breaking long sentences into meaningful words for easier analysis. It also lets you **track word usage, patterns, and themes across centuries**, revealing hidden connections in literature and historical documents.

Week 6 10 Oct

Part-of-Speech Tagging

- After learning part-of-speech tagging, you can **analyze the grammar and style of historical texts**, comparing how authors used verbs, nouns, or adjectives over time. It also allows you to **detect linguistic patterns, sentence structures, or shifts in language use**, uncovering subtle trends in literature and historical writing.

Week 7 17 Oct

Named Entity Recognition

- After learning named entity recognition, you can **automatically extract people, places, and events from historical texts**, turning unstructured documents into structured data. This makes it possible to **map historical networks, trace movements of figures, or explore connections across time and geography** in a insightful way.

Week 8 24 Oct

Text Classification

- After learning text classification, you can **automatically sort large collections of historical or literary texts by topic, genre, or sentiment**, saving tons of manual work. You could even **discover hidden trends or patterns**, like how themes of love, war, or philosophy evolve over centuries.

Week 9 31 Oct

Text Clustering

- After learning text clustering, you can **create beautiful cluster maps or visualizations** that show how texts relate to each other. It's good to **see patterns, similarities, or unexpected groupings** across historical documents or literary works in an intuitive, visual way.

Week 10 7 Nov

Using GitHub and Markdown

- Using GitHub and Markdown, you can **organize and present your digital humanities projects online**, making them **ready for publication or academic sharing**. It's also fun to **showcase your work interactively and let others explore your projects easily**.

Week 11 14 Nov

Group Presentations

Week 12 21 Nov

Group Presentations

Week 13 28 Nov

Discussion Class (topics will be decided based on the class situation)

Readings

** Specific chapters will be assigned by the instructor before class*

- Gold, Matthew K. Debates in the digital humanities. University of Minnesota Press, 2012.
- Nugues, Pierre M. Python for Natural Language Processing: Programming with NumPy, scikit-learn, Keras, and PyTorch. Springer, 2024.