CHIN 303/DATA 340: Hacking Chinese Studies

Instructor: Paul Vierthaler (pavierthaler@wm.edu)
Office: Washington Hall 234
Office Hours: Tuesdays 10-12 and by appointment

Washington Hall 308
Monday and Wednesday 17:00-18:20

This course is an introduction to leveraging the processing power of modern computers to study Chinese culture. This course will introduce you to a variety of newly developed digital tools, algorithms, and datasets that allow us to pursue new insights into traditional Chinese literature and culture. You will engage with new scholarship being published in the rapidly expanding field of the digital humanities and learn how to create digital research projects from scratch. You will be introduced to the basics of text mining, network analysis, mapping, and digital exhibition creation, among other things. We will draw examples from the imperial Chinese tradition and cover the challenges and rewards of working with Chinese language materials using computer systems originally designed for western languages (the lessons we learn will be applicable to people working in other non-Latinate languages like Japanese, Hebrew, Arabic, and Sanskrit). While some of the materials we will learn to analyze are in Chinese, we will also work with English language datasets and literature in translation. As such, the class will be open and accessible to students who do not know any Chinese. **No technical skills or programming experience required. There are no prerequisites for this class.**

While this course is focused on China, it will give students a new set of digital tools that can be used across a wide variety of fields, so **students from all disciplines are encouraged to join**.

In this class you will learn the basics of programming in **Python**. We will begin each session by introducing a programming concept and by the end of the semester you will have enough command of the language that you will be able to write small programs that can clean data, perform various analyses, and visualize it. You should bring your laptop to class. If you do not have access to a laptop, please reach out to me (pavierthaler@wm.edu) and we will make sure you are still able to participate.

Format and Expectations

This class will include lectures, interactive discussion, and group work. Active participation is expected, so please come to class prepared after having read the materials.

Attendance Policy:

I will not be taking attendance, but active participation in class is required and part of your grade. If you miss class consistently, you will do poorly in participation. Many of the concepts we will cover are cumulative and it will be easy to fall behind. If you miss a session, be sure to get notes from your fellow classmates.

Assignments
There will be eight assignments throughout the semester that are based on materials we cover in class (this may be writing a small computer script to perform an analysis, creating a visualization from a dataset, or writing reflective essays on the material). **These assignments will all be posted on Blackboard one week before they are due and will be announced in class**.

Final Project/Paper
The students will finish the course with a final project of 2,750 words (+/- 10 percent, this should be roughly ten to twelve double-spaced pages). This can be a paper, a website, or a multimedia project with content roughly equivalent to the paper. Every paper should include at least two data visualizations. **The fifth assignment will be a proposal for your project. If you are doing a creative project/one that does not easily convert to a word count, include a section describing what you think constitutes a successful project (which will, after any modifications I think are necessary, be used as the rubric for grading your project)** In the last week of class, students will present their projects (5 to 7 minutes). The final paper is due at **10 pm on May 6th**. I am happy to comment on drafts, but please send them to me at least one week before the paper is due, otherwise I will be unable to provide comments.

Late assignments and papers: will be penalized by a 10% reduction for each 24-hour period it is late. After one calendar week, the assignment will not be accepted.

Please ensure that your papers/projects/digital files can be opened and read properly on a Mac OS X, Windows 10, or Ubuntu computer with standard software (Acrobat, Word, WordPad, etc). If you have any concerns, contact me ahead of time. Corrupted or un-openable files will be considered late.

Writing Expectations
While content is the most important component of your work, a portion of your grade on each assignment will be for style and understandability. There is a writing resource center at William & Mary. If you find yourself struggling, please feel free to avail yourself of their services! Alternatively, come speak with me and I will do what I can to help.

A Word About Plagiarism
You must document all of your source material. If you take any text from somebody else, you must make it clear the text is being quoted and where the text comes from. You must also cite any sources from which you obtain numbers, ideas, or other material. If you have any questions about what does or does not constitute plagiarism, ask! Plagiarism is a serious offense and will not be treated lightly. Fortunately, it is also easy to avoid and if you are the least bit careful about giving credit where credit is due you should not run into any problems (thanks to Alfred E Guy, Jr. for this statement).

You are encouraged to use the internet to help you with the programming assignments. If you find a solution to a programming issue online at website like StackOverflow, cite the source and **explain why it works for the task at hand**.

Grading Rubric:

Attendance and Participation: 10 percent
Assignments: 55 percent
Final presentation: 10 percent
Final paper: 25 percent

GRADING SCALE FOR FINAL GRADES

A 93 – 100	A- 90 – 92.9	
B+ 87 – 89.9	B 83 – 86.9	B- 80 – 82.9
C+ 77 – 79.9	C 73 – 76.9	C- 70 – 72.9
D+ 67 – 69.9	D 63 – 66.9	D- 60 – 62.9
F <60		

Required Books

All materials for this class are freely available online and open source.

Please take advantage of this online tutorial series, which covers all of the Python concepts we discuss in this class (disclaimer: I made these. They are not required but might be useful): https://www.youtube.com/playlist?list=PL6kqrM2i6BPIpEF5yHPNkYhjHm-FYWh17

Readings will be distributed through blackboard.

Accommodations:

Student Accessibility Services: William & Mary accommodates students with disabilities in accordance with federal laws and university policy. Any student who feels they may need an accommodation based on the impact of a learning, psychiatric, physical, or chronic health diagnosis should contact Student Accessibility Services staff at 757-221-2512 or at sas@wm.edu to determine if accommodations are warranted and to obtain an official letter of accommodation. For more information, please see www.wm.edu/sas.

Schedule (subject to change):

Week 1: Class Introduction

Jan 22: What are the digital humanities and why should we care?
Why should we build our own tools?
Syllabus

Intro to Programming and Text analysis

Week 2: The Basics

Jan 27: Anne Burdick,et.al, "From Humanities to Digital Humanities," 1-26.

https://www.dropbox.com/s/zcfhiphslciqe2k/9248.pdf?dl=1

Paul Vierthaler "The State of Chinese Digital Humanities in the West" *Draft, not for circulation*

Technical: The Command Line, Strings, Integers

Jan 29: Digital Chinese Texts: A brief history of Chinese corpora

CText, CBETA/SAT, Kanseki
Read: https://ctext.org/introduction; http://blog.kanripo.org/en1.html

Laura McGrath, "More Specific, More Complex"
http://post45.research.yale.edu/2019/05/more-specific-more-complex/

Technical: Floats, Math, Lists

Week 3: Processing (Chinese) Information

Feb 3: Computing in Chinese in an ASCII world (or, what ARE character encodings?)

Technical: Booleans, Loops, Files

Feb 5: Data structures and models: How we store information and why it influences how we process it.

Technical: Dictionaries

Week 4: Natural Language Processing: Basic linguistic analysis

Feb 10: Errors, Functions, and NLTK intro
"Language Processing and Python"
https://www.nltk.org/book/ch01.html

Technical: NLTK/CLTK, Loading Corpora

Feb 12: Chinese Word Segmentation

Read: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

Sproat et al "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese" https://www.aclweb.org/anthology/J96-3004.pdf (feel free to skim the math-heavy sections).

Technical: NLTK and Chinese in practice

Week 5

Feb 17: Data Extraction from Text
Doug Knox, “Understanding Regular Expressions”
https://programminghistorian.org/en/lessons/understanding-regular-expressions

Technical: Regular Expressions

Feb 19: From Unstructured to Structured:
Transforming the *Annotated Catalog of the Complete Library of the Four Treasuries* 四庫全書總目提要

https://zh.wikisource.org/zh-hant/四庫全書總目提要

Week 6:
Feb 24: Matplotlib, Seaborn, and the basics of Plotting with Python

Feb 26: BeautifulSoup, robots.txt, why we scrape, and what are the ethics?

Wendy Hsu, “Digital Ethnography: Toward Augmented Empricism”
http://journalofdigitalhumanities.org/3-1/digital-ethnography-toward-augmented-empiricism-by-wendy-hsu/

Technical: Web scraping the *Annotated Catalog*

Week 7:

March 2: Stylometry pt 1
Cristof Schoch: “Principal Component Analysis for Literary Genre Stylistics”
https://dragonfly.hypotheses.org/472

Peter Turney and Patrick Pantel, “From Frequency to Meaning: Vector Space Models of Semantics,” Journal of Artificial Intelligence Research 37 (2010): 141-188

Technical: Hierarchical Cluster Analysis

March 4: Stylometry pt 2

Video on Stylometry (13 minutes):
https://www.youtube.com/watch?v=jZ532ucT6Ik&list=PL6kqrM2i6BPLakYAnvoXE6pUJ9T1qGHXd

Paul Vierthaler, “Fiction and Stylistic Gradience Late Imperial Chinese Literature,” https://culturalanalytics.org/2016/05/fiction-and-history-polarity-and-stylistic-gradience-in-late-imperial-chinese-literature/

Technical: Principal Component Analysis

Week 8: **Spring Break (March 7 – 15)**

Week 9: Machine learning and the humanities

March 16: Topic Modelling
Ted Underwood, “Topic-modelling made just simple enough:” http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Allen, et. al., “Topic Modeling the Handian Ancient Classics” http://culturalanalytics.org/2017/10/topic-modeling-the-han-dian-ancient-classics-%E6%B1%89%E5%85%B8%E5%8F%A4%E7%B1%8D/

Technical: Topic Modelling

March 18: Word Embedding Models
Sarah Connel, “Word Embeddings are the New Topic Models” https://web.northeastern.edu/nulab/word-embedding-model/

Lisa Rhodey, “Topic Modeling and Figurative Language” http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

Technical: Word embedding models and Gensim

Week 10: Networks

March 23: Network Analysis pt 1

Scott Weingart, “Demystifying Networks, Parts I & II”: http://journalofdigitalhumanities.org/1-1/demystifying-networks-by-scott-weingart

Technical: Network analysis in Python

March 25: Network Analysis pt 2

Chen Song, "Governing a Multicentered Empire: Prefects and Their Networks in the 1040s and 1210s.” In State Power in China, 900-1325, edited by Patricia

Buckley Ebrey and Paul Jakov Smith. Seattle, WA: University of Washington Press, 2016.

Technical: Visualizing Networks

Week 11: Mapping

March 30: Mapping I

Patricial Murrieta-Flores, Christopher Elliott Donaldson, and Ian Norman Gregory. "GIS and literary history: advancing digital humanities research through the spatial analysis of historical travel writing and topographical literature." http://www.digitalhumanities.org/dhq/vol/11/1/000283/000283.html

Chinese Historical GIS: http://chgis.fas.harvard.edu/pages/intro/, http://chgis.fas.harvard.edu/pages/history/

Technical: Tools for Mapping in Python

April 1: Mapping II

JavaScript and Interactive Map-making

Week 12: Markup

April 6: IIIF, Images, and Image Markup

Tina Lu and Mick Hunter, *The Ten Thousand Rooms Project*
Explore: https://tenthousandrooms.yale.edu/

April 8: Text Markup, TEI, and the MARKUS project

Week 13: Databases

April 13: Database Creation and Chinese Studies

Relational databases, Graph databases, and Structuring Data Meaningfully

April 15: The China Biographical Database

Week 14:

April 20: Online publication and the Public Humanities

Robyn Schroeder, “What is public humanities?”
https://dayofph.wordpress.com/what-ispublic-humanities

Tom Mullaney, ed, *The Chinese Deathscape*
Explore: http://chinesedeathscape.org/

Technical: Creating an online exhibition

April 22: Class wrap-up and Discussion

Week 15: Presentations

April 27: Presentations Day 1

April 29: Presentations Day 2

Your final project is due at 10 pm on May 6th! Submit it on Blackboard!