From Text to GIS: A Reflective Journey of Scholars Using GenAI and DH Tools for Spatial Analysis

Link : https://drive.google.com/drive/folders/1Enu3FuQHWwlmHxgQM05cmV766YnmAzY3?usp=sharing

Introduction

My traditional way of reading novels like "The Scholars" has always been "close reading" - paying attention to the specific plot and characters line by line. However, this task presents a new challenge: how to systematically analyze the spatial distribution of activities across multiple chapters. This requires a shift towards "remote reading", a digital humanities (DH) approach that analyzes literature by aggregating and analyzing large amounts of data. In this article, I will reflect on my workflow of converting unstructured text into GIS visualizations using Python, Streamlit, and GenAI, highlighting how these tools bridge the gap between literary studies and data science.

Methodology and workflow

My workflow begins with data collection. I copied the original text files of the relevant chapters from Ctext.org. However, these files are unstructured plain text, making manual counting of position frequencies cumbersome and error-prone. To solve this problem, I used Python and the pandas library for data processing.

I wrote a Python script to traverse text files, calculate the occurrence frequency of specific cities (Nanjing, Suzhou, Hangzhou, Beijing, Yangzhou, Jinan, Huzhou), and export the context to an Excel file.

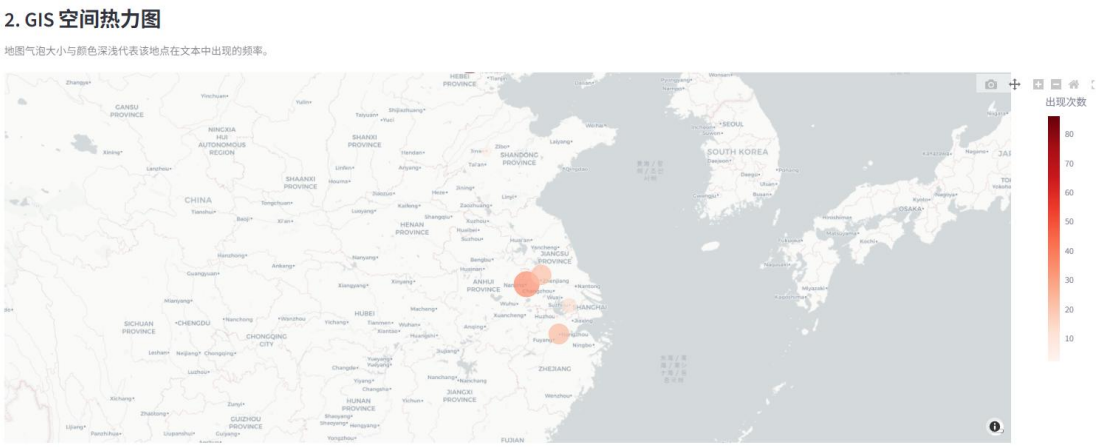
```python
target_locations = {
    "南京": ["南京", "金陵", "秦淮"],
    "苏州": ["苏州", "姑苏", "吴门"],
    "杭州": ["杭州", "西湖", "武林", "钱塘"],
    "北京": ["北京", "京师", "京", "长安", "都门", "帝京"],  # 增加了大量别名
    "扬州": ["扬州", "维扬", "广陵"],
    "济南": ["济南", "山东", "大明湖", "历下"],  # 用山东代指济南区域
    "湖州": ["湖州", "吴兴"]
}

# 用来存放结果
summary_data = []
context_data = []
```

Python script for data extraction, with an alias mapping dictionary

"Aha!" Critical moment: Solving the Problem of Hidden data In my research, a critical moment occurred during the initial data analysis process. The result of my first attempt was that the frequency in Beijing and Jinan was "zero". This contradicts my knowledge because I know the character will go to the capital. After carefully reading the text, I realized that the novel used historical terms: Beijing was called "the Capital" or "Chang 'an", while Jinan was often hinted at as "Shandong".

This understanding highlights a key limitation of the DH tool: the tool is text-based; They lack cultural background. To solve this problem, I improved my algorithm by introducing the "Alias Mapping" dictionary (as shown in Figure 1). I programmed this tool to recognize "Beijing Capital" as "Beijing City". This adjustment successfully revealed the hidden data and proved that domain knowledge is indispensable when guiding computing tools.

After cleaning up the data, I visualized the results using Streamlit and Plotly. I didn't use static charts but built an interactive web site. This enables me to draw positions on a geospatial map, where the size of the bubbles corresponds to the frequency of the mention.



The interactive dashboard visualizing the spatial distribution

Reflections on tools

Reflecting on this experience, I have gained several insights into the role of GenAI and DH tools in cultural studies:

First of all, GenAI has lowered the technical barriers. Without GenAI providing code structure and debugging assistance (for example, fixing file permission errors), I might have spent several days setting up the environment. It enables me to focus on the logic of analysis (for example, defining aliases) rather than the syntax of the programming language.

Second, visualization changes the perspective. The heat map generated by Streamlit provides a direct visual confirmation of the novel's "southern-centered" narrative. Seeing the dense red dots around Nanjing and Yangzhou, in contrast to the sparse ones in the north, provides a macro-level insight that is more difficult to grasp when reading chapter by chapter.

Finally, people are indispensable in the cycle. The mistake of "Beijing/Jingshi" tells me that data is never neutral or readily available. It requires human explanations. If I blindly believe in the first output of the machine, I will draw the wrong conclusion that Beijing is unimportant in these chapters. The power of DH does not lie in replacing researchers, but in enhancing their ability to discover patterns, provided that researchers guide the tools with cultural expertise.

In conclusion, this practical operation assignment is not merely a coding exercise; This is a course on interdisciplinary research. By combining Python's computing power with critical literary analysis. Tools offer "what it is" and "where it is", but my reflection and careful reading provide "why".

DIGITAL HUMANISTIC
ANALYSIS AND REFLECTION

From Political Center to Cultural Sanctuary

LIN YANG 25124747G
CHC5904

RESEARCH DIRECTION

RESEARCH QUESTIONS

01

METHODOLOGY & TOOLS

CTEXT.ORG

1 →

PYTHON

2 →

```python
import nltk
from nltk import ngrams
from collections import Counter  # 确保导入Counter（之前代码可能已有，重
nltk.download('punkt_tab')  # 已下载过可忽略，首次运行需执行

# 1. 收集所有章节的完整文本
all_text = ""
input_dir = "data"  # 确保路径和你原有代码一致
for filename in os.listdir(input_dir):
    if filename.lower().endswith(".txt"):
        file_path = os.path.join(input_dir, filename)
        try:
            with open(file_path, 'r', encoding='utf-8') as f:
                text = f.read()
        except:
            with open(file_path, 'r', encoding='gbk') as f:
                text = f.read()
        all_text += text + " "  # 拼接所有文本

# 2. 分词并生成2元组N-gram
tokens = nltk.word_tokenize(all_text)  # 分词
bigrams = list(ngrams(tokens, 2))  # 2元组搭配

# 3. 只关注"行者"和"三藏"
target_aliases = {'行者', '三藏'}
```

STREAMLIT

3

02

WORKFLOW

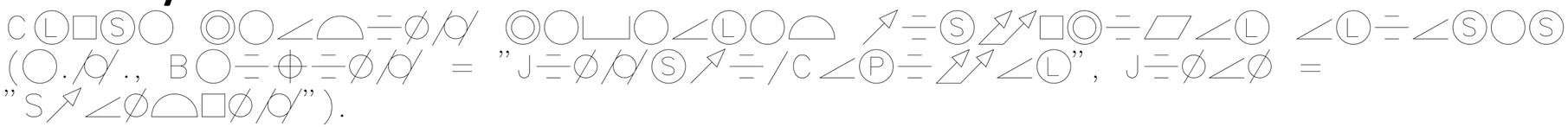
```
target_locations = {
    "南京": ["南京", "金陵", "秦淮"],
    "苏州": ["苏州", "姑苏", "吴门"],
    "杭州": ["杭州", "西湖", "武林", "钱塘"],
    "北京": ["北京", "京师", "京", "长安", "都门", "帝京"],  # 增加了大量别名
    "扬州": ["扬州", "维扬", "广陵"],
    "济南": ["济南", "山东", "大明湖", "历下"],  # 用山东代指济南区域
    "湖州": ["湖州", "吴兴"]
}

# 用来存放结果
summary_data = []
context_data = []
```

03

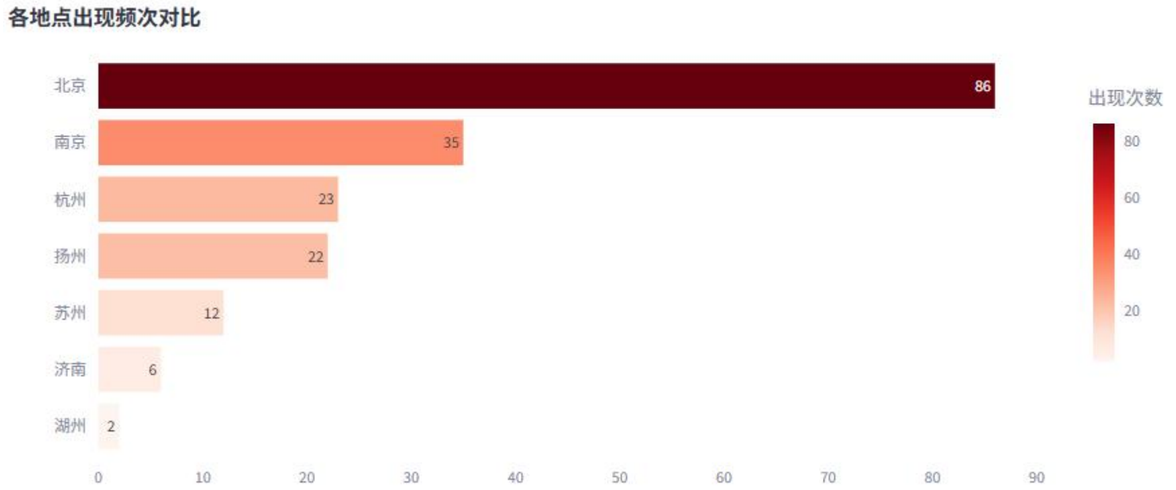
Data Processing Challenge

Problem:

Discovery:



Solution:



Observation:
The frequency chart shows a decisive dominance of Southern cities.

Data Point:
Nanjing and **Yangzhou** appear significantly more often than **Beijing**.

Interpretation:
The novel moves away from the center of political power (North) to the centers of culture and commerce (South).

04

Quantitative Findings

GIS Visualization (Spatial Analysis)

- **Visualization:** An interactive heatmap generated by Plotly/Streamlit.
- **Analysis:** The map visualizes the character's mobility. The "Hot Zone" is clearly concentrated in the Yangtze River Delta (Jiangnan), reflecting the vibrant scholar culture in this region.

06

]:.....老爷现住在承恩寺。差人说，请少爷在家里，邓老爷自己上门来请。”杜少卿道：“既如此说，我不走前门家去了。你快叫一只船，我从河房栏杆上上去。”当下小厮在下浮桥雇了一只凉篷，杜少
。忙取一件旧衣服，一顶旧帽子，穿戴起来，拿手帕包了头，睡在床上，叫小厮：“你向那差人说，我得了暴病，请邓老爷不用来，我病好了，慢慢来谢邓老爷。”小厮打发差人去了。娘子笑道：“
官，你为甚么妆病不去？”杜少卿道：“你好呆！放著南京这样好顽的所在，留著我在家，春天秋天，同你出去看花吃酒，好不快活。为甚么要送我到京里去？假使连你也带往京里，京里又冷，你
阵风吹得冻死了，也不好。还是不去的妥当。”小厮进来说：“邓老爷来了，坐在河房里，定要会少爷。”杜少卿叫两个小厮搀扶著，做个十分有病的模样，路也走不全，出来拜谢知县；拜在地下，就
知县慌忙扶了起来，坐下就道：“朝廷大典，李大人专要借光，不想先生病得狼狈至此。不知几时可以勉强就道？”杜少卿道：“治晚不幸大病，生死难保，这事断不能了。总求老父台代我恳辞。”袖
呈子来递与知县。知县看这般光景，不好久坐，说道：“弟且别了先生，恐怕劳神。这事，弟也只得备文书详覆上去，看大人意思何如。”杜少卿道：“极蒙台爱，恕治晚不能躬送了。”知县作别上轿而
了文书，说：“杜生委系患病，不能就道。”申详了李大人。恰好李大人也调了福建巡抚，这事就.....

Case Study 1 Nanjing

(The Cultural Ideal)

Theme:

The City of "Ritual and Music" (礼乐).

Evidence:

Textual analysis reveals activities such as the "Grand Assembly at Mochou Lake" (莫愁湖大会).

Insight:

Nanjing serves as the spiritual home for scholars, representing high culture and intellectual freedom away from officialdom.

07

[28.txt]: ……奶奶不曾听见怎的，你怎么又做这件事？”季苇萧指著对联与他看道：“你不见‘才子佳人信有之’？我们风流人物，只要才子佳人会合，一房两房，何足为奇！”鲍廷玺道：“这也罢了。你这些费用是那里来的？”季苇萧道：“我一到扬州，荀年伯就送了我一百二十两银子，又把我在瓜洲管关税。只怕还要在这里过几年，所以又娶一个亲。姑老爷，你几时回南京去？”鲍廷玺道：“姑爷，不瞒你说，我在苏州去投奔一个亲戚投不著，来到这里，而今并没有盘缠回南京。”季苇萧道：“这个容易。我如今送几钱银子与姑老爷做盘费，还要托姑老爷带一个书子到南京去。” 6 Jump to dictionary 季苇萧扬… : 正说著，只见那辛先生、金先生，和一个道士，又有一个人，一齐来吵房。季苇萧让了进去，新房里吵了一会，出来坐下。辛先生指著这两位向季苇萧道：“这位道友尊姓来，号霞士，也是我们扬州诗人。这位是芜湖郭铁笔先生，镌的图书最妙。今日也趁著喜事来奉访。”季苇萧问了二位的下处，说道：“即日来答拜。”辛先生和金先生道：“这位令亲鲍老爹，前日听说尊府是南京的，却几时回南京去？”季苇萧道：“也就在这一两日间。”那两位先生道：“这等，我们不能同行了。我们同在这个俗地方，人不知道敬重，将来也要到南京去。”说了一会话，四人作别去了。鲍廷玺问道：“姑爷，你带书子到南京与那一位朋友？”季苇萧道：“他也是我们安庆人，也姓季，叫作季恬逸，和我同姓不宗。前……

Case Study 2 Yangzhou

(The Secular World)

Theme:

The City of "Commerce and Pragmatism".

Evidence:

Context extraction highlights salt merchants, matrilocal marriages (入赘), and social networking.

Insight:

Yangzhou represents the intersection of literary skill and commercial money. It is more secular and grounded compared to Nanjing.

CONCLUSION

08

Revisiting Research Questions:

Beijing is the "Mental Center" (Most mentioned, feared, respected).

Jiangnan is the "Physical Center" (Where life actually happens).

Reflection on Tools:

GenAI & Python: Essential for solving the "Alias Problem" (Beijing vs. Jingshi). Without this, the analysis would have been factually wrong.

Streamlit Map: Provided a "Distant Reading" view that highlights the geographic span of the scholars' network.

Final Thought:

Digital tools help us see the structure (Frequency/Map), while Close Reading helps us understand the sentiment (Why they prefer Nanjing over Beijing).


THANK YOU FOR
LESTENING!