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.

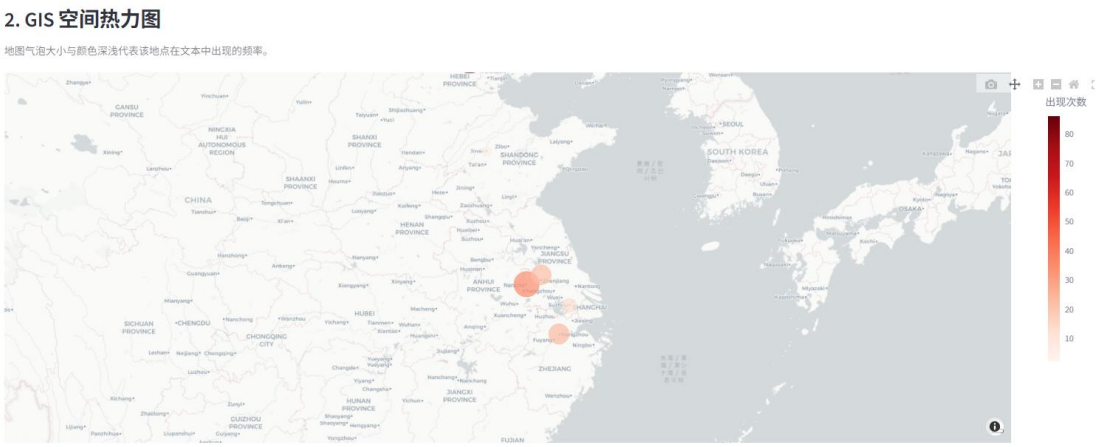
```python
target_locations = {
    "南京": ["南京", "金陵", "秦淮"],
    "苏州": ["苏州", "姑苏", "吴门"],
    "杭州": ["杭州", "西湖", "武林", "钱塘"],
    "北京": ["北京", "京师", "京", "长安", "都门", "帝京"],  # 增加了大量别名
    "扬州": ["扬州", "维扬", "广陵"],
    "济南": ["济南", "山东", "大明湖", "历下"],  # 用山东代指济南区域
    "湖州": ["湖州", "吴兴"]
}

# 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".