**Final Project**

Yilin Du

Computational Literacy

20.12.2024

# The usage of color vocabulary in *Dream of the Red Chamber*

## Introduction

*The Dream of the Red Chamber* (*Hongloumeng*, also known as *The Story of the Stone*) is one of the Four Great Classical Novels of Chinese literature. It describes the rise and fall of a grand family. The use of color-related vocabulary in *Dream of the Red Chamber* is unparalleled in ancient Chinese novels and reflects the broad and profound concepts of traditional Chinese color symbolism. As a novel that delves into human emotions, vibrant colors run throughout its vivid narrative. The author, Cao Xueqin, came from a family where four generations held the prestigious position of Jiangning Weaving Commissioner, a role connected to mastery of color and textile knowledge. This expertise is evident in his detailed use of color throughout the novel.

Research on the use of color in Dream of the Red Chamber is relatively scarce, and most studies focus on the symbolic meanings of color from a literary perspective. Few have conducted comprehensive statistical analysis or explored color vocabulary across the entire text.

This project aims to explore two main questions:

Can digital tools effectively process ancient Chinese novels like Dream of the Red Chamber?

What are the characteristics of color vocabulary usage in the novel? What are the main types and their usage contexts?

## Data

The data used in this study is the complete text of Dream of the Red Chamber in txt

format (https://github.com/hankinghu/literature-books/blob/master), with approx. 960,000 words. The version is the one edited and annotated by the Red Chamber Dream Institute of the China Academy of Art, published by the People's Literature Publishing House. The author, Cao Xueqin (1715-1763), wrote the novel between the early and the 30th year of the Qianlong reign (1736-1765).

The color terms are categorized into seven basic colors: red, green, yellow, blue, purple, black, and white. Under each category, there are several specific words representing variations of that color. The list includes:

Red-related words: 红 (red), 丹 (cinnabar), 朱 (vermilion), 赤 (scarlet), 绛 (carmine), 胭脂 (rouge), 茜 (madder), 猩 (crimson), 血色 (blood-red), 紫绛 (purple-red), 玫瑰 (rose), 绒 (velvet), 春色 (spring color), 荔色 (lychee color), 杨妃色 (Yang Fei color).

Green-related words: 绿 (green), 翠 (emerald), 碧 (jade).

Yellow-related words: 黄 (yellow), 金 (gold), 杏 (apricot), 秋香色 (autumn fragrance), 松花色 (pine flower), 土色 (earth color), 蜜合色 (honeycomb color).

Blue-related words: 青 (cyan), 月白 (moon-white), 蓝 (blue), 雨过天晴（after-rain sky color).

Purple-related words: 紫 (purple), 藕合色 (lotus-pink), 茄色 (eggplant color), 酱色 (chestnut).

Black-related words: 黑 (black), 缁 (dark), 玄 (deep), 墨 (ink), 皂 (dark brown).

White-related words: 白 (white), 银 (silver).

**Process**

I first downloaded the txt file of the complete text of *Dream of the Red Chamber* (https://github.com/hankinghu/literature-books/blob/master) and removed irrelevant content such as publishing information. I then imported the text into VS Code and used Python for further text processing.

The first step involved extracting color words and their contextual usage. I defined seven dictionaries, each corresponding to a color category, and each dictionary included specific vocabulary for that color. I also defined a stopword list, which included common stopwords and some proper nouns, such as names of characters that contain

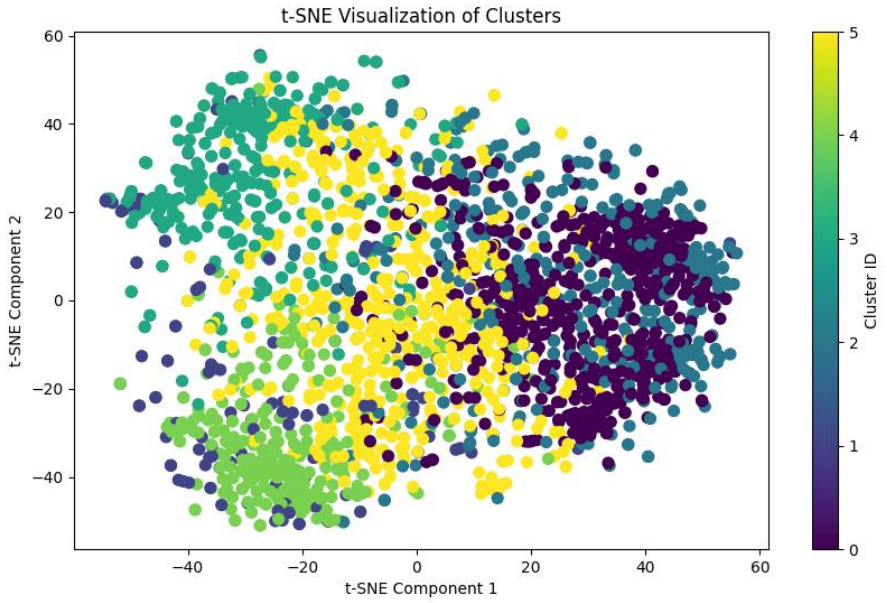
color words but do not refer to actual color usage. I removed punctuation from the text and determined that the context length would be ten characters before and after each color word (the maximum length). I then extracted the color words and their context within one category and saved the results in separate CSV files. So I get seven CSV files: red_context.csv, green_context.csv, yellow_context.csv, blue_context.csv, purple_context.csv, black_context.csv, white_context.csv.

In the process of extracting and processing these color words, I used to attempt tokenization. Tokenizing Chinese text is challenging because there are no spaces between words, and it requires semantic and contextual judgment, which digital tools often struggle with. Since *Dream of the Red Chamber* is written in a transitional language style between Classical Chinese and modern vernacular Chinese, tokenization becomes even more complex. I tried three existing tools for segmenting Chinese text: Jieba (https://github.com/fxsjy/jieba), Jiayan (https://github.com/jiaeyan/Jiayan), and HanLP (https://github.com/hankcs/HanLP), but Jiayan's installer had some problems and I couldn't use it in Python, and Jieba as well as HanLP didn't perform well in this kind of half-classic, half-vernacular text segmentation, often failing to identify words containing color words and splitting them into separate parts with only one Chinese character in each part. Thus, I opted to extract the context around each color word instead of performing tokenization.

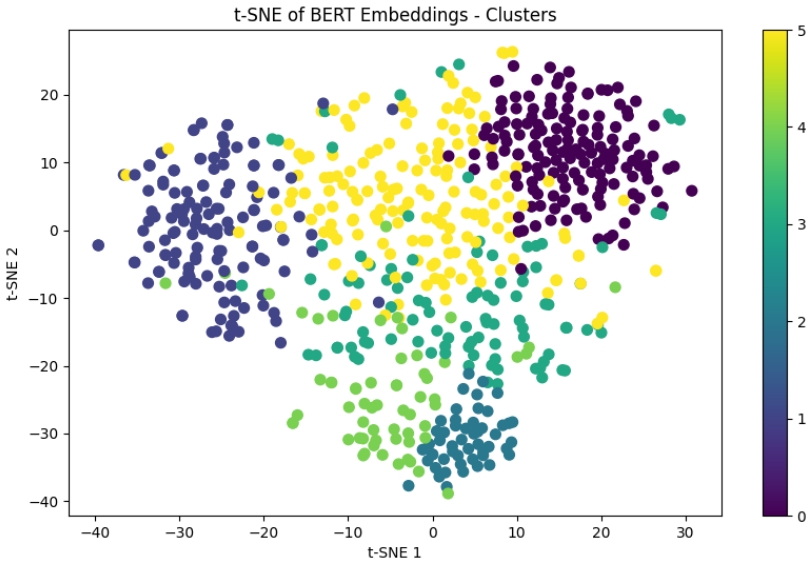
After extracting the context, I performed clustering to analyze the categorize and usage of color words (as seen in *all_categories.py*). I first tried to merge the seven CSV files for an overall clustering of all the color words. I got a 1616 line total CSV file and used the transformers library from Hugging Face to load Chinese related BERT splitters with models to generate embedding vectors for the text. The embedding vectors of each context were obtained by the BERT model, which transformed each text into a high-dimensional vector representation. And then, the similarity between the texts is calculated by cosine similarity. The texts were clustered using K-means clustering algorithm and different number of clusters were tried. And after that, Silhouette Score is used in order to evaluate the clustering effect with different number of clusters. I tried six cluster numbers for range(3, 9) and the best performer was 5 with a Silhouette Score

of 0.21634 followed by 6 with a Silhouette Score of 0.17569. However, the results of this clustering were not ideal, with no clusters exceeding a 0.25 Silhouette Score.

When clustering into six groups, t-SNE visualizations is shown below:



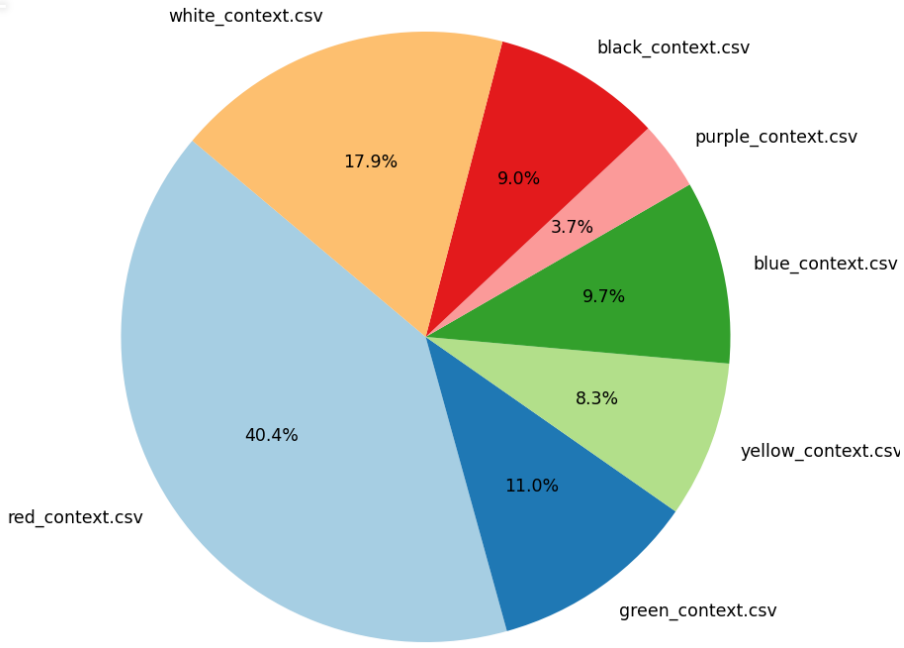

The clustering results are not significant, there are no clear boundaries between the categories and there is a lot of overlap. Extracting representative words for each cluster and looking at the first 50 also shows that the second and third categories are particularly confusing and often misclassified. I then changed the length of the contexts to 4, 6, 8 and 12 characters when extracting the color word contexts in the first step, and the final clusters I obtained were not significantly better. At this point, I believed that the semantic confusion caused by the variety of color words was likely contributing to the issues, so I next tried further to go for clustering within a specific color category, starting with the most frequent one: red. The number of clusters remains 6 (as seen in *red_related_words.py*). The t-SNE visualization is as follows:




On the basis of the above treatment, further analysis can be carried out on the frequency of use of color words and the situations in which they are used.

**Analysis**

After extracting the seven categories of color words and the specific contexts in which they appeared, I obtained seven CSV files with 663 entries in the red category, 178 entries in the green category, 134 entries in the yellow category, 157 entries in the blue category, 59 entries in the purple category, 290 entries in the white category and 145 entries in the black category.

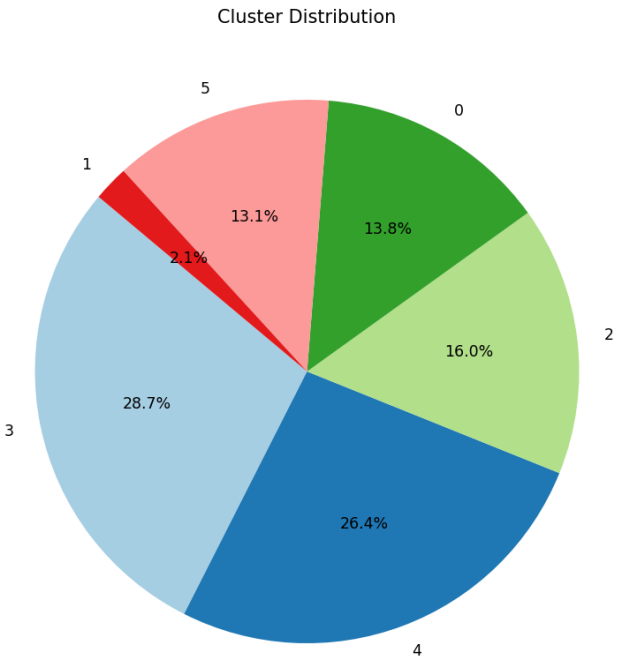


It can be concluded that red is the dominant color in *Dream of the Red Chamber*, and red-related color terms form the largest and most frequent category in the text. The scope of red's usage is also the broadest, encompassing not only specific colors used to describe characters' clothing, appearance, and objects but also numerous abstract color terms conveying subjective emotions. The main objects associated with the color red are clothing, fabrics, utensils, and plants, with the largest number of instances occurring in the description of clothing. In ancient Chinese color tradition, red was a color of high rank, symbolizing status and power. Only high-ranking officials and aristocrats were permitted to use red on a large scale, and it also had the function of warding off evil and bringing good fortune. Characters associated with red are typically the wealthy and powerful figures in the Jia family within this book, such as Grandmother Jia, Baoyu, and Wang Xifeng. Moreover, the word "red" itself appears in the title *Dream of the Red Chamber*, referring to beautiful young women and the fleeting nature of youth. The depiction of red as the most supreme and precious color represents another form of the "red reverence" consciousness. The reason behind this reverence is the comparison of red to a daughter: on the surface, it expresses love for the color red, but in reality, it reflects the author's respect for women. Additionally, the term "red dust" (红尘), which has Buddhist origins, is often used in everyday language to signify the transient, fleeting nature of life.

The second most frequently appearing color is white. On one hand, white represents "silver", the color of ancient Chinese currency. Silver items were considered valuable, and they were commonly used by noble families, aligning with the status of the Jia family depicted in *Dream of the Red Chamber*. On the other hand, white is also the color of snow, symbolizing the inevitable end of all things and the return to dust. The term "white vastness" (白茫茫), with its connotations of the impermanence of life, reflects the Daoist philosophical concept of the inevitable decline of worldly affairs, a major theme in *Dream of the Red Chamber*. The novel revolves around the fall of a wealthy family, and this unavoidable decay is expressed through the symbolism of white. Additionally, "snow white" is used to describe the beauty of women, often highlighting their pure and virtuous nature.

Next come green, blue, black, yellow, and purple. Green is primarily used to describe plants and other natural scenery. A significant portion of the novel is dedicated to describing the Grand View Garden (大观园), the residence of Jia Baoyu and the noble young ladies. The garden is filled with trees, and green frequently appears in the descriptions of these natural surroundings. Blue is associated with the color of the sky and carries connotations of life, but in this novel, it is mainly used to describe clothing. Specifically, the colors stone blue (石青) and moon white (月白) were popular among the Qing nobility. Black is mostly used to describe objects, such as the ink used for writing, as well as clothing and shoes. Yellow is the color of gold and represents a higher value than silver. In ancient China, yellow was also a symbol of imperial authority, and only the royal family could wear yellow clothing. In *Dream of the Red Chamber*, yellow is predominantly used to describe natural scenery and objects made from gold. Purple, the least frequently used color, was complex to produce and difficult to obtain directly from nature, making it a rare color. When purple does appear in the novel, it usually signifies wealth and luxury in the Jia family, as seen in Jia Baoyu's purple clothing and accessories.

The code used above effectively implements the extraction and storage of color terms. Next comes the clustering process. After carefully reviewing the color phrases in the CSV file, I categorized them into six usage scenarios myself: characters, plants and nature, clothing and accessories, architecture, utensils, and lyrical/philosophical meanings. Based on this categorization, I set up six clusters for the subsequent clustering code. I used the BERT model to generate embedding vectors for each color phrase and calculated the similarity between texts using cosine similarity. Then, I applied the K-means clustering algorithm to group the texts. The distribution of the final results is as follows:




The number of color phrases in each cluster is:

Cluster 3 463

Cluster 4 426

Cluster 2 259

Cluster 0 223

Cluster 5 211

Cluster 1 34

By extracting the most representative terms from each cluster and displaying the top 50, it became clear that:

Cluster 0 corresponds to clothing and accessories.

Cluster 1 corresponds to architecture.

Cluster 2 corresponds to characters.

Cluster 3 contains both clothing and plants/nature.

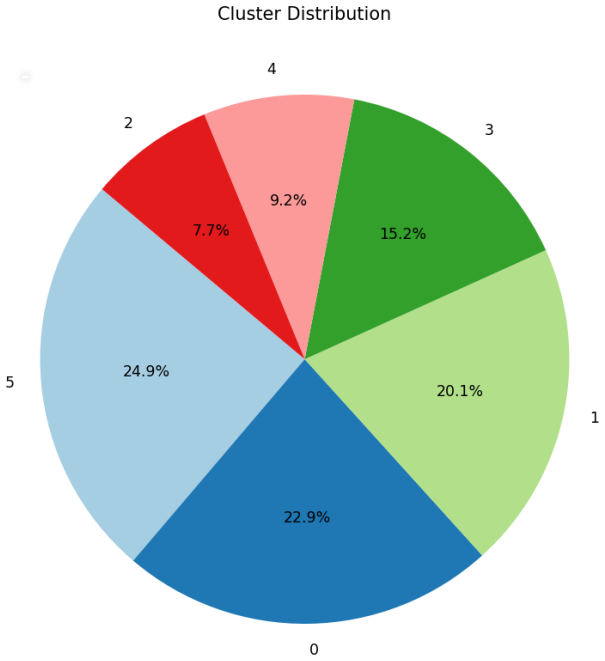
Cluster 4 contains characters, clothing, and utensils.

Cluster 5 corresponds to lyrical/philosophical themes.

The use of color terms in the architecture category was the least frequent. In contrast, color terms used to describe characters, clothing, and lyrical themes were more evenly distributed. Utensils and plants/nature, however, were not clearly identified as standalone categories, as they were mixed with other categories, making it difficult to

distinguish them.

Next, I decided to focus on classifying the use of a single color category. I chose red, the most frequent and significant color in Dream of the Red Chamber, and continued to use the BERT model with the K-means clustering algorithm. The clustering results for the red color category are as follows:




The number of color phrases in each cluster is:

Cluster 5 165

Cluster 0 152

Cluster 1 133

Cluster 3 101

Cluster 4 61

Cluster 2 51

The specific categorization is as follows:

Cluster 0 corresponds to characters.

Cluster 1 corresponds to clothing and accessories.

Cluster 2 corresponds to lyrical/philosophical themes.

Cluster 3 includes utensils and architecture.

Cluster 4 corresponds to plants and nature.

Cluster 5 includes both clothing and architecture.

Although there is still some overlap between the categories, the boundaries have become clearer compared to the previous overall clustering. More specifically, the color terms related to red primarily fall into two categories: describing characters (脸红 (blushing), 哭红了眼(eyes reddened from crying)) and describing the color of clothing and accessories (桃红袄 (peach red coat), 硬红镶金坠子(bright red gold-embellished pendant)). There are fewer occurrences of red being used to describe plants, utensils, or architecture, like to describe the color of flowers (红海棠(red begonia), 红梅(red plum)), or to describe the color of a building's facade (朱楼(vermilion building)). Additionally, red is also used to convey emotion and philosophy, such as lamenting the fleeting nature of youth and beauty, with 红颜(red face) symbolizing women: 一朝春尽红颜老，花落人亡两不知（When spring ends, the red face grows old, while flowers fall and people perish, both unaware); or reflecting on the transience of life and the uncertainty of fate: 昨夜朱楼梦，今宵水国吟 （Last night in the red tower dreamt, tonight in the water kingdom sang).

In conclusion, the use of color in Dream of the Red Chamber is intricate and symbolic, reflecting both the material wealth of the Jia family and the novel's deeper philosophical themes, such as the impermanence of life and the fleeting nature of beauty and youth. The prominence of red and white, in particular, serves to highlight the complex interplay of cultural, social, and emotional meanings associated with color in this seminal work of Chinese literature. Meanwhile, red, as the dominant color in the title Dream of the Red Chamber, not only represents the tangible beauty of life and youth but also alludes to the underlying illusory, transient, and uncertain nature of that beauty and youth. The use of red in the novel points to the vibrant, ephemeral moments of life but also carries an implicit reminder of life's impermanence.

**Problems and bias**

One of the primary objectives of this project is to evaluate whether existing digital tools can effectively process classical Chinese texts. This uncertainty has inevitably led to various challenges throughout the analysis process.

The nature of the data itself presents significant obstacles for natural language processing tools. Unlike English or other alphabet-based languages, Chinese lacks spaces between words, requiring contextual and semantic understanding to determine word boundaries. This is a fundamental challenge for digital tools. The language style of *Dream of the Red Chamber*, situated at the transition from classical to modern Chinese, further complicates matters. The linguistic features of this transitional period are not adequately captured by existing tools, making their application less effective in this context. Moreover, the method I currently use to identify and extract color terms and their contexts tends to overlook nuanced elements of the text, such as idiomatic expressions, historical references, and wordplay. Extracting entire sentences for analysis risks losing the subtleties of the original language, thereby complicating subsequent clustering tasks. Incorporating a specialized dictionary or manually expanding the algorithm's vocabulary to include commonly used terms within the novel's context might help address these issues and improve accuracy in identifying color terms and their applications.

The clustering process introduces further challenges, especially given the reliance on unsupervised algorithms such as K-means. These methods often fail to account for the intricate nature of human language, particularly in cases where category boundaries are ambiguous. Without domain-specific supervision, the identification of some contexts for color terms is inaccurate, and the overlapping nature of the categories adds to this complexity. Additionally, literature allows for ambiguity and polysemy, which is especially true for *Dream of the Red Chamber*. For example, the color terms used in the novel—particularly those associated with "red"—carry rich cultural and symbolic meanings. Phrases like "red face" or "red tower" encapsulate layers of significance that modern algorithms struggle to grasp. As a result, automated tools often prioritize surface-level interpretations, neglecting deeper emotional or philosophical associations.

A potential solution to this problem would be to adopt supervised machine learning techniques. By training models on a manually annotated dataset, the algorithm could better differentiate between specific categories, capturing more nuanced distinctions. Although annotating the entire dataset would be labor-intensive, a smaller subset could

be labeled to train the model, with the remaining data used to refine its predictions. This semi-supervised approach would balance the need for precision with practical constraints. However, even manual annotation introduces its own set of challenges. Decisions made during the annotation process could reflect personal biases, particularly when determining the context for color usage. Such subjectivity may inadvertently skew the data toward specific interpretations or themes.

Another area for improvement lies in the tools used for processing the text. While I have experimented with widely used tools such as Jieba, Jiayan, and HanLP, I have not yet explored resources specifically designed for classical Chinese, such as ANTT (Ancient Chinese Text Tool) or software developed for analyzing classical poetry. These tools, being more attuned to the linguistic and cultural context of classical Chinese texts, may provide better results for *Dream of the Red Chamber*. Additionally, employing domain-specific language models or embeddings trained on classical Chinese could enhance the analysis, especially in recognizing idiomatic expressions and culturally significant terms.