Natural Language Processing for Character Analysis

ZHAOYI CHENG (21-740-279)

1.Introduction

This project is dedicated to employing advanced Natural Language Processing (NLP) techniques to meticulously extract character descriptions and develop character networks from select literary texts. By integrating the capabilities of the BookNLP pipeline, this analysis will focus on identifying and categorizing characters, their attributes, interactions, and the intricate relationships they share within the narrative framework.

The methodology is structured as follows: Initially, I will carefully select literary works that are most suitable for this analysis and configure the BookNLP pipeline to prepare the text data for preprocessing. Subsequently, I will apply NLP techniques to systematically extract character names, and their descriptive attributes, and document their interactions. Finally, the project will culminate in the creation of a detailed character network that visually represents the relationships and dynamics among the characters, derived from the analyzed data.

2.Dataset

The books used in the project as datasets are all from Gutenberg (cite the website). Since most natural language processing models currently use English data for training, and most of the books on Gutenberg are in English, I also chose English books as our dataset. First, I typed ‘fiction’

into the search box. Then, fiction books sorted by download volume appear on the results page. I randomly selected 6 books from the homepage and downloaded them in the old ePub format.

The first book is 'Alice in Wonderland'(alice or alice wonderland in the report). This book is one of the top three most downloaded on the site, making it popular. Besides the main character, most of the main "characters" are small animals. This means that the descriptions of their appearances might differ from those of human characters.

The second book is 'The Great Gatsby'(Gatsby in the report). This book also ranks high in download volume, making it one of the popular books. A characteristic of this book is that it contains many characters with complex interrelationships, and there are many descriptions of parties in the book. This means that the characters' clothing styles might be very similar, which will help us observe whether generative models produce different images for similar texts.

The third book is 'A Room with a View'("Roomtour" in the report). The characters in this book involve different ages, personalities, and social roles. This means we can extract keywords from this book that describe the appearances of contemporaneous characters, and these keywords have meanings that differ without being anachronistic.

The fourth book is 'The Blue Castle'("blue castle" in the report), where the protagonist experiences significant emotional and residential changes. This means that the words used to describe the appearance of the same character in the book might vary greatly.

The fifth book is 'Enchanted April'("enchanted april in the report"). The protagonists are four distinctly different women from England who come

together for a vacation in Italy, where they reconcile with their pasts and confront a life filled with hope and love. The characters' mentalities and situations undergo noticeable changes, which means their appearances would also change correspondingly.

The sixth book is 'Little Women'(“little woman” in the report). This book mainly revolves around the journey of four sisters from childhood to womanhood. This means that descriptions of the appearance changes of the same character at different ages might appear in the book.

Table 1.

Book	wrapper	intro	Number of chapters	content	others
The blue castle	y	y	45	y	\
A roomtour	y	y	20	y	\
Alice wonderland	y	y	12	y	\
Enchanted april	y	y	22	y	\
Gatsby	y	y	9	y	“Once again to Zelda”
Little woman	y	y	48	y	list of illustration

3. Methodology

3.1 TORI Pipeline

TORI is a content generation pipeline designed for preprocessing EPUB files. It includes various components such as the EPUB pipeline, image processing pipeline, text processing pipeline, and sound pipeline.

In this project, we primarily utilize certain functions from the EPUB pipeline:

- read_book: Reads an EPUB book from a specified file.
- chapter_to_str: Converts a chapter object into a string.
- is_valid: Validates the structure of an EPUB book, ensuring the presence of the table of contents, book spine, and stylesheets, and confirming that the chapters are not empty.
- is_wrapper: Determines if the item is the Gutenberg cover wrapper.
- is_gutenberg_intro: Determines if a part of the content is the Gutenberg introduction.

By using these functions within the TORI pipeline, we can remove non-novel components from the book, such as the cover and introductions. The filtered content is then compiled into a new text for subsequent NLP analysis.

3.2 BookNLP

In this project, BookNLP, a natural language processing tool tailored for analyzing novels and other extensive English texts, was utilized to facilitate automated text file analysis. This initial step involved extracting structured data about entities, events, quotes, and coreferences within the narrative. Consequently, the output files containing detailed information on entities and tokens are prepared for subsequent analyses.

Subsequently, the extracted book data was employed to gather basic information about the characters, such as names, frequency of appearance, gender, and interaction behaviors. It was assumed that characters who

appear more frequently are the main characters of the book. A character list was created for these prominent figures.

To analyze character interactions, nested loops were implemented to calculate the interactions between all pairs of these characters within the same sentence. This approach allowed for the detailed examination of character dynamics and relationships as portrayed in the narrative.

4 Results

4.1. BookNLP vs TORI&BookNLP

4.1.1. Book Spine

I tested the TORI pipeline on six books. The TORI pipeline works well if the books follow a regular format, which includes a wrapper, introduction, and table of contents. It accurately recognizes the content of each chapter. It correctly matches the content with the chapter numbers, as seen in "The Blue Castle," "Enchanted April," "Alice's Adventures in Wonderland," and "A Room with a View."

However, issues arise when the format includes additional elements. In "The Great Gatsby," the content of the first chapter is categorized as part of the header. Chapters four, six, and nine have their content merged with the preceding chapters. Consequently, during the filtering process, any text categorized as part of the header is removed, resulting in incomplete content. This misclassification affects character feature extraction, leading to missing or duplicated features due to incorrect chapter associations.

A similar issue occurs with "Little Women," where parts of the novel's content are not correctly classified. This misclassification results in missing text, causing similar problems.

Observations reveal that correctly classified novels share a common format, with no additional structure beyond the cover, introduction, and table of contents. In contrast, novels with incorrect classifications, like "Little Women," have unique formats. For instance, each chapter in "Little Women" contains sub-chapters, with each sub-chapter having a complete or incomplete sentence as a title. These sub-chapter titles resemble the main text but have their own independent table of contents, potentially causing the misclassification.

Books	results
The Blue castle	All contents are readable correctly
Gatsby	Chapter 1, 2, 4, 6, and 9 are not divided correctly. Chapter 4 is included in chapter 3.
enchanted_ april	All contents are readable correctly
alice_wonderland	All contents are readable correctly
room_tour	All contents are readable correctly
little_woman	part of the contents can be readable correctly

Table 2. Chapter division by TORI

4.1.2. TORI performance

I will use data processed by the TORI pipeline and unprocessed original data as inputs to BookNLP. I will analyze the .book files from the output. By comparing the mention counts of the main characters in the book, if the mention count decreases, we will consider that TORI effectively pre-processes the data, simplifies the text structure, or reduces redundant information, thereby focusing more accurately on the most important content in the book.

From table 3, in most cases, the TORI treatment resulted in a decrease in the number of mentions, which may indicate the pipeline's ability to reduce redundancy or simplify the text by focusing on what is important.

There were some exceptions, such as ‘Gatsby’ and ‘Mrs. Fisher’ in ‘Enchanted April,’ where the number of mentions increased, suggesting that the TORI treatment may have made certain key characters more prominent or emphasized.

As for ‘Room Tour’, the TORI treatment did not affect the number of mentions, implying that the text may have been optimized or that the effect of TORI was minimal.

Books	Character 1	Character 2	Character 3
The Blue Castle(original)	4096(Valancy)	676(Barney)	343(Uncle Benjamin)
The Blue Castle (TORI)	3978(Valancy)	618(Barney)	322(Roaring Abel)
Gatsby(original)	960(Gatsby)	711(Daisy)	580(Tom)
Gatsby (TORI)	1013(Gatsby)	660(Daisy)	507(Tom)
Enchanted April (original)	1660(Mrs. Fisher)	1246(Mrs. Wilkins)	868(Mr. Wilkins)
Enchanted April (TORI)	2440(Mrs. Fisher)	1145(Mrs. Wilkins)	858(Mr. Wilkins)
Alice wonderland(original)	1823(Alice)	102(the Mock Turtle)	100 (the Hatter)
Alice wonderland (TORI)	1671(Alice)	102(the Hatter)	93(the Mock Turtle)
room tour(original)	2331(Lucy)	793(Cecil)	664(George)
room tour (TORI)	2331(Lucy)	793(Cecil)	664(George)
Little woman(original)	5310(Jo)	2952(Laurie)	2607(Amy)
Little woman (TORI)	5073(Jo)	2445(Meg)	2257(Amy)

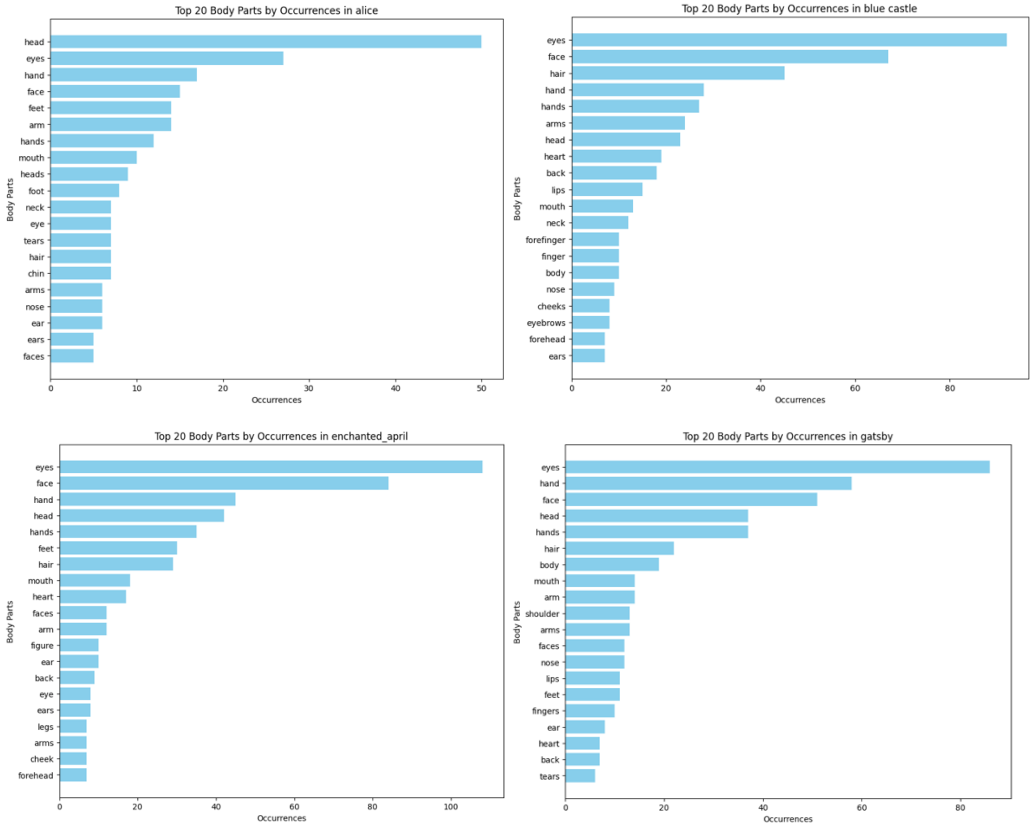
Table 3. A comparison of the number of times characters are mentioned before and after processing with the TORI pipeline

4.2. Character description words extraction

4.2.1. Wordnet

First, I analyzed the frequency of words related to physical appearance or body parts mentioned in the books. I identified and saved the top 20 most frequently mentioned words for each book, and then visualized this data. The results show that terms such as "eyes," "face," and "head" consistently appear with high frequency across the books. However, words like "nose," "mouth," and "hair" are less frequently mentioned and do not always make it into the top 20 in some of the books.

To enhance the likelihood of capturing more descriptive words, I utilized WordNet to identify synonyms, hypernyms, and hyponyms related to these terms, and saved this extended list of related words in a text file.



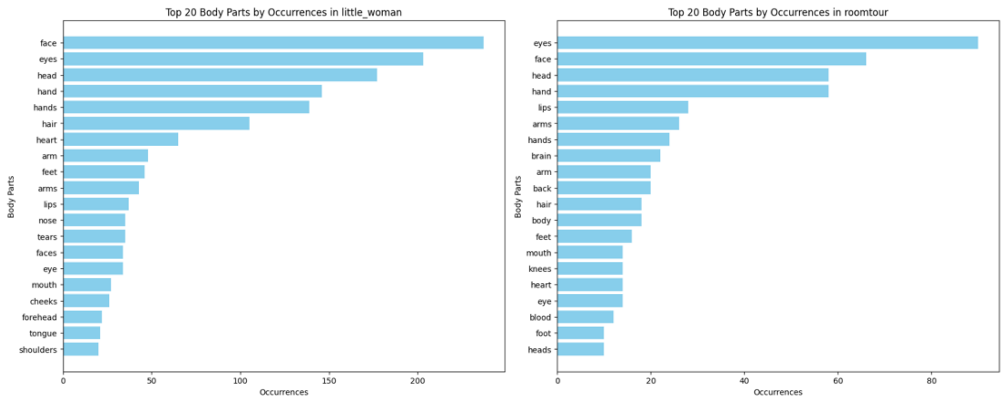


Fig 1. The top 20 appearance words in each book

4.2.2. Information extraction

Based on the previously calculated frequency of character names, we selected the top three characters from each book for descriptive phrase extraction. Across the 18 characters analyzed, there are 10 females, 8 males, and one anthropomorphized animal—a turtle—reflecting the varied narratives of each book. The extracted phrases were organized in chronological order according to the chapters, allowing for easy comparison of whether characters undergo any changes in their appearance over the course of the story. The extracted effective phrases primarily focus on frequently mentioned body parts such as eyes, face, and hands. However, some phrases, like "the look," are included due to hypernyms and other higher-level terms. Despite setting up filters to remove prefixes like "the/his/her/a/an," some phrases with these prefixes still appeared. Notably, while "hair"， ”face”, ”eyes” are frequently mentioned words, detailed descriptions are limited, often constrained to general terms like "long hair."

One potential issue could stem from inaccuracies in coreference resolution. Descriptive phrases associated with pronouns like "she" or "he" may be incorrectly attributed or entirely overlooked. Although attempts were made to use NeuralCoref and BERT for coreference resolution, these efforts were unsuccessful due to pip installation failures with NeuralCoref. The

plugin's stringent version requirements for different dependencies made it difficult to install successfully, even when attempting to manually configure the environment.

4.3. Social Network

In the final step of the analysis, I visualized the social networks of the top 10 most frequently mentioned characters in each book. The edges represent the number of interactions between characters—specifically, how often they appear together in the same sentence or context. However, in the case of Alice in Wonderland, this method failed to accurately capture the interaction counts, likely due to the presence of numerous anthropomorphized characters. Additionally, for Little Women, there was an issue when reading the quotes file and converting it to a data frame, which resulted in mismatches with the entities and subsequently made it impossible to generate the social network.

Taking the social network graph of The Blue Castle as an example, Valancy and Barney, who occupy central positions, are among the most frequently mentioned characters in the book. The social network clearly shows that they have the most significant interactions with other characters. Another interesting observation is that Roaring, despite being the third most frequently mentioned character, is positioned on the periphery of the network. This could be related to the storyline or the way the character is portrayed in the book.

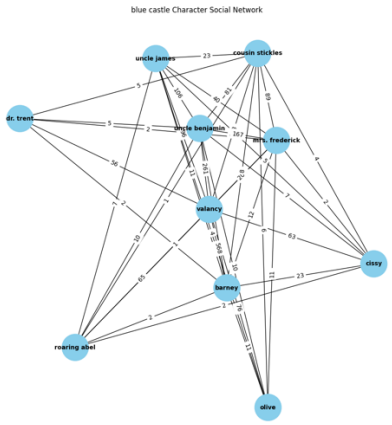


Fig 2. Example figure of social network (the Blue Castle)

5. Discussion

5.1. Chapter Division

When using the TORI pipeline to perform chapter division for The Great Gatsby and Little Women, we encountered instances where multiple chapters were merged into a single chapter. Specifically, in The Great Gatsby, every two chapters were identified as one. In Little Women, some chapters were incorrectly merged with the preceding chapter. Comparing the structure of these two books with others, we noticed that both contain sections resembling an introduction before the main text. Therefore, we speculate that these chapter division errors may be caused by the unique structure of the books.

5.2. Issues about NeuralCoref installation

When working on tasks like information extraction and social network analysis, the most effective solutions consistently highlight the importance of NeuralCoref. However, when I attempted to install NeuralCoref via pip following the README guide, the process repeatedly failed. The errors were related to compatibility issues with other plugin versions, and even

manually installing the specified versions did not resolve the issues. Later, I attempted to set up a virtual environment and install NeuralCoref within it, but this also failed. Despite these challenges, I believe that integrating NeuralCoref would significantly enhance the value of the project's output. Accurate coreference resolution is crucial for precisely extracting character information and exploring character relationships.

5.3. Issues about information extraction

Due to the failure of installing NeuralCoref, I attempted to extract more appearance-related phrases using a basic coreference method. I used the vocabulary list obtained from WordNet as the search base and focused on extracting phrases for the top three characters. In this approach, I assumed that descriptive phrases in sentences containing pronouns like "she" or "he" referred to the nearest character's name, thereby implementing a simple form of coreference resolution. However, this approach is evidently too simplistic, leading to many instances where the meaning of the text, which requires contextual understanding, is incorrectly assigned to the nearest character, resulting in erroneous coreference.

6. Conclusion

This project focuses on extracting descriptive phrases related to character appearances from books in the Gutenberg collection, applying NLP techniques within the realm of information extraction. During the initial phase of data processing, I employed the TORI pipeline and, through comparative analysis, confirmed its effectiveness in simplifying and structuring textual information. For character analysis, the main characters

were identified based on the frequency of mentions, allowing for a quantitative approach to determining the protagonists. Using a basic coreference method, I extracted appearance-related phrases for the top three characters, organizing these phrases in chronological order according to the chapters. Additionally, I used interaction network visualization to illustrate the strength of interactions between characters, further validating that the main characters identified by their high frequency of mentions were accurately chosen.

7. Further extension

In the future, more accurate coreference algorithms could be applied to extract character appearance phrases and reorganize the data. The extracted phrases could then be used as inputs for generating character avatars or images that closely align with the narrative's descriptions, thereby enhancing the textual analysis with a visual dimension.

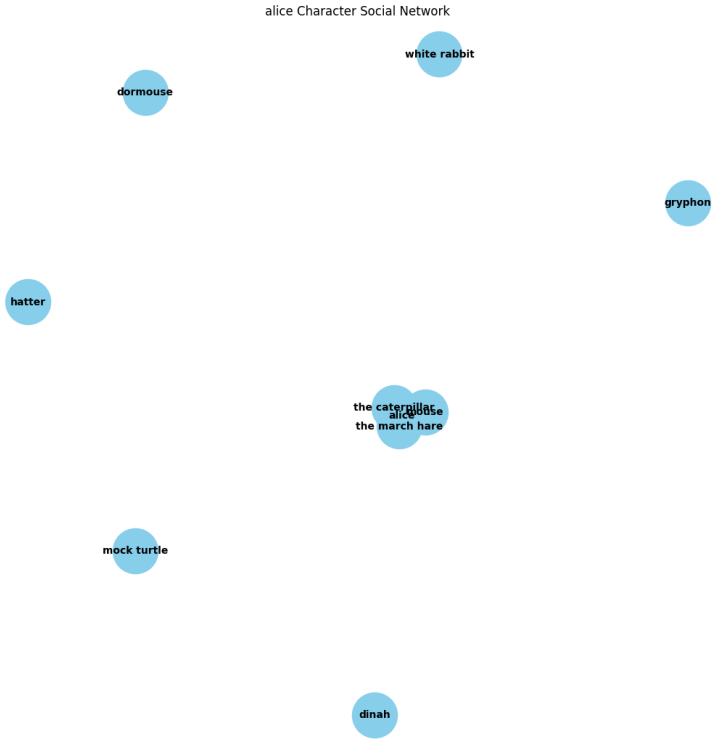
Reference

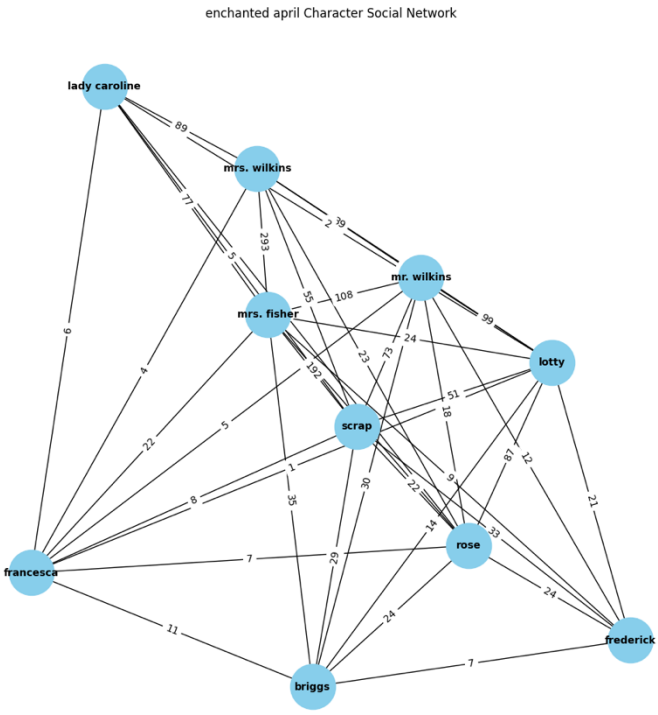
https://github.com/booknlp/booknlp/blob/main/LICENSE

List of AI tools

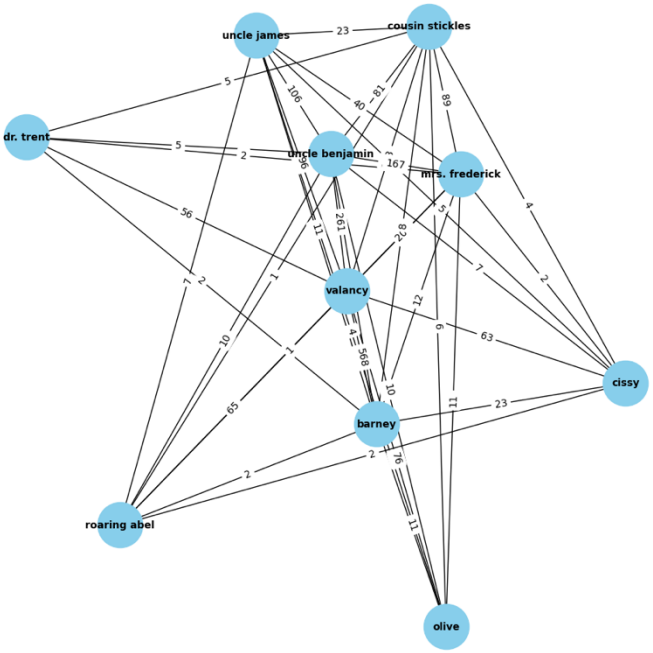
1. ChatGPT – translating some words and sometimes debugging
2. DeepL – translating some words

Appendix A. Social Network

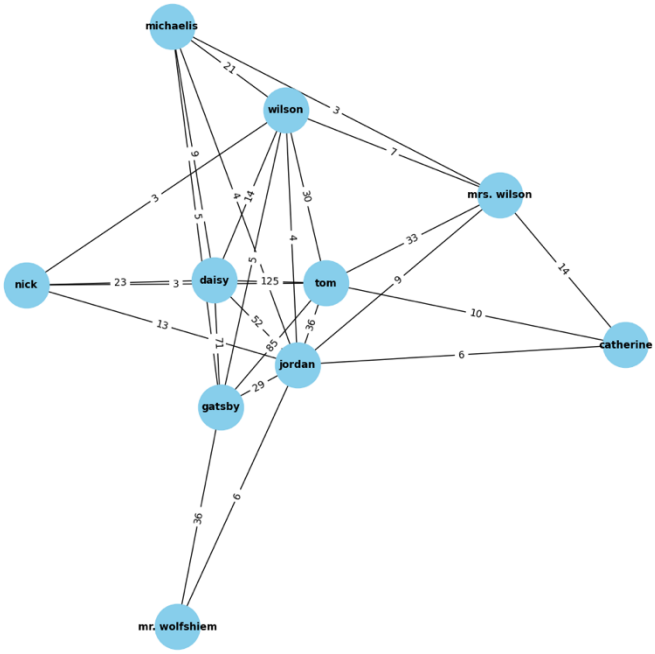


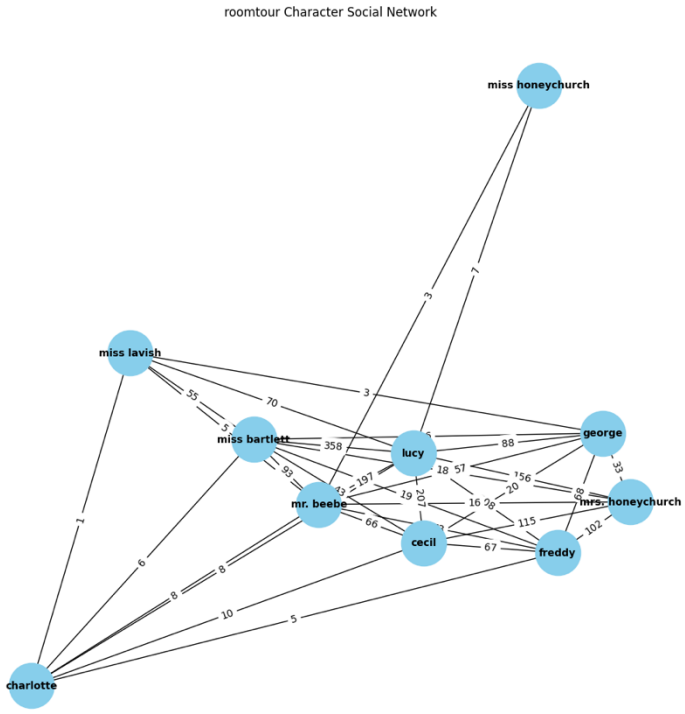


blue castle Character Social Network



gatsby Character Social Network




roomtour Character Social Network
miss honeychurch
miss lavish
miss bartlett
lucy
george
mr. beebe
mrs. honeychurch
cecil
freddy
charlotte
3
7
3
70
55
358
88
57
156
33
93
197
18
19
20
16
26
20
66
115
102
67
1
6
8
8
10
5