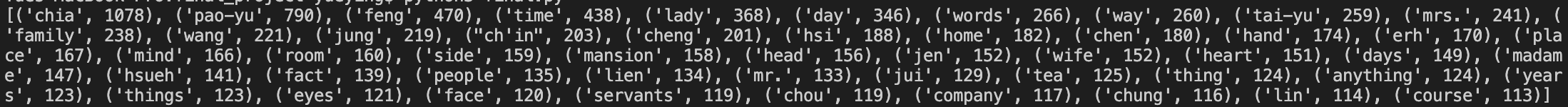
DRC Textual Analysis Output

* Most common collocates

Text

Description automatically generated

* Frequency distribution of most common nouns (top 50)



* It’s worth noting that “pao-ch’ai” is not even in this list, while both “pao-yu” and “tai-yu” are very high in the list. Also, even when “pao-ch’ai” is vaguely associated, the noun that occurred was “heueh” instead of her surname. She was referred more to her last name than surname, underlining the intimacy between “pao-yu” and “tai-yu” instead of “pao-ch’ai”. From the high occurrence of “feng” and “lady” (might be referring to lady Chia”, we can even conclude that “pao-ch’ai” is not among the most important female characters in this book, which is somewhat on the contrary as what the general public might think.
* Get words used in a similar context as “chia” (limit to 10) 贾



* From here, we see that the Chia is often associated with female names or female pronouns. The “feng” and “chou”, two names, that show up in the similar context as “chia” are females closely related with the “chia” family → in fact, run the “chia” family, but they do not have “chia” in their last name. The other thing that’s worth noting is “chin”真, indicating that truth and false always come together.
* Get words used in a similar context as “heueh” (limit to 10) 薛



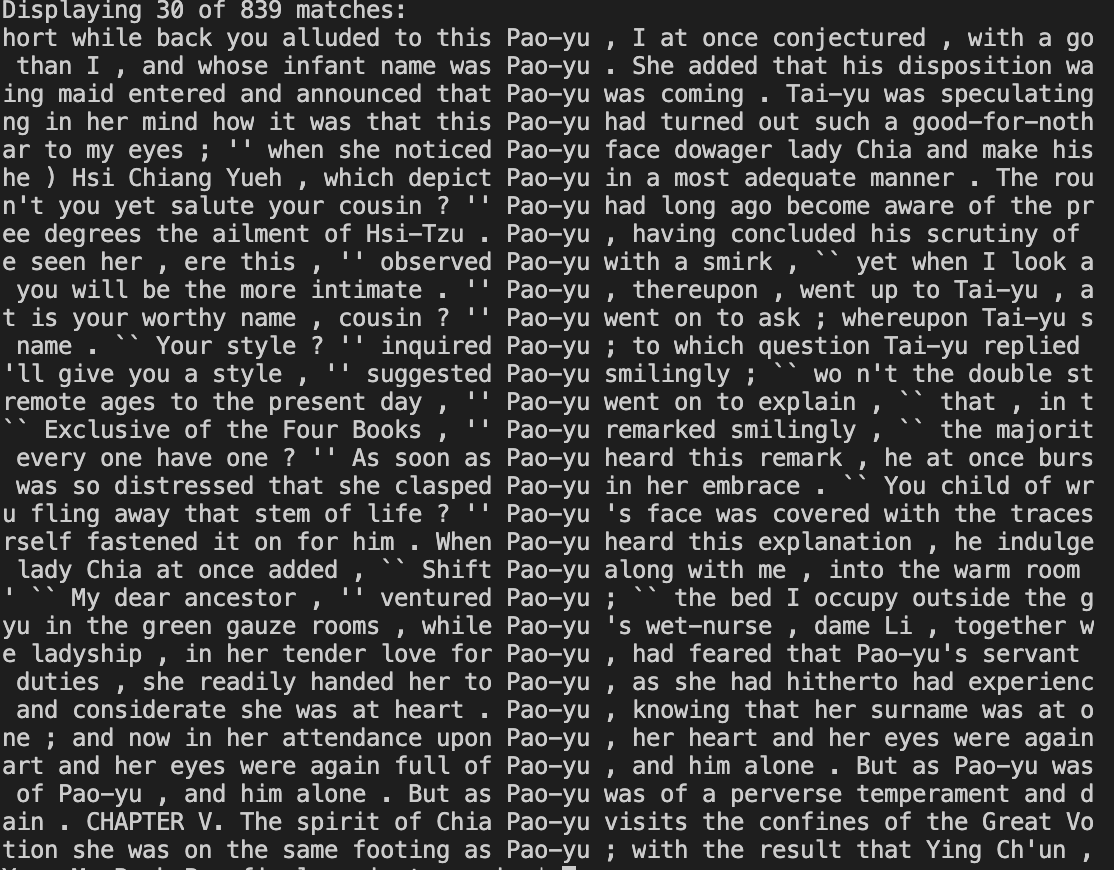
* We see that when薛is mentioned, it is always mentioned in a similar setting as林and贾combined. It seems that she is naturally the third-wheeler in林黛玉and贾宝玉’s relationship. “they” occurs very frequently in a similar text as薛, indicating that薛is mentioned as a singular unit while林and贾are mentioned as a joint unit together.
* Get words used in a similar context as “lady” (limit to 10)



* In the Chia family, ladies are put in a similar context as master, as gentlemen, as father (the traditional dominant figure at home), and as someone with high intelligence (“mind”). This is a book of feminism and awakening, written in Qing Dynasty.
* Get words used in a similar context as “maid” (limit to 10)



* The “maids” are not used in a context of slavery, instead, the maids here (in the Chia family) are a part of the family. They provide accompany to their masters. They can have their own purpose, life goal, and they are also well-education to converse and even create poems with their masters.
* Concordances of “pao-yu” → listing every instance of each word with its immediate context



* We can hardly find a time when Pao-yu is not mentioned with a female pronoun, either “she” or “her” or a reference of a female name. This is a book of just the females and female power, instead of saying that only men can save the family. In fact, the corruption of the family comes from males. Pao-yu, as the future of the family, has all the potentials to maintain the prosperous status of his family. He also has a natural inclination for that (contrary to what the large audience thought – they believed that Pao-yu didn’t study and only wanted to do female things). But the problem occurs that he is not fully a female. He is still a male and being treated as a male, with all male expectations. Thus, he cannot fully immerse himself to the female world. While his sisters are able to handle family matters, he could not. It seems that his male self was hindering himself and his potential to rescue the family. The females in the family are all pure, intelligent, well-rounded and smart, yet the males in this households are all corrupt and dirty. Pao-yu, as someone in the middle, struggles in understanding his identity and purpose of life. Although he didn’t get trapped to that male corrupt world, he did not inherit all the qualities that the females in his household had. This constant struggle finally led him to ordination, the final epiphany of him becoming a monk.
* Moreover, from the direct text, we can see that “Pao-yu:” is often mentioned with “Tai-yu”. At least in the top 30 lines of occurrences, we did not find a single one occurrence that happened with “Pao-ch’ai”.

**Compare Similarity between Book I and Book II**

* Book I Chapter I



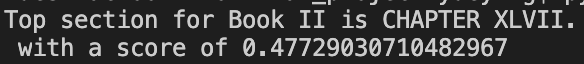
* Book I Chapter II



* Book I Chapter III



* Book I Chapter IV



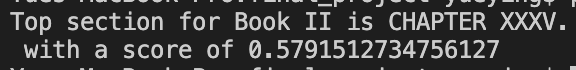
* Book I Chapter V



* Book I Chapter VI



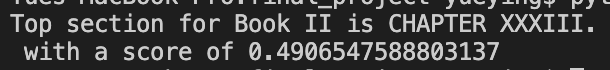
* Book I Chapter VII



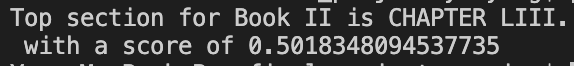
* Book I Chapter VIII



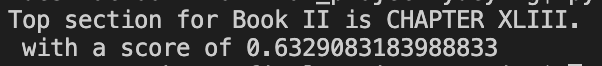
* Book I Chapter IX



* Book I Chapter X



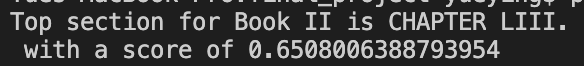
* Book I Chapter XI



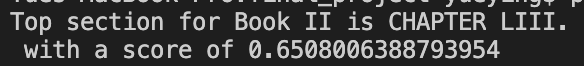
* Book I Chapter XII



* Book I Chapter XIII



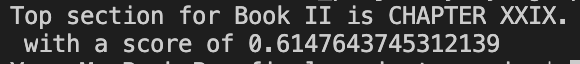
* Book I Chapter XIV



* Book I Chapter XV



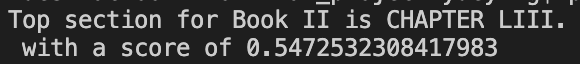
* Book I Chapter XVI



* Book I Chapter XVII



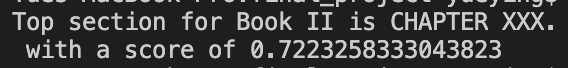
* Book I Chapter XVIII



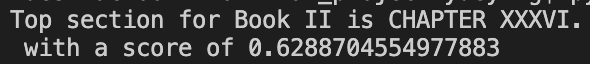
* Book I Chapter XIX



* Book I Chapter XX



* Book I Chapter XXI



* Book I Chapter XXII



* Book I Chapter XXIII



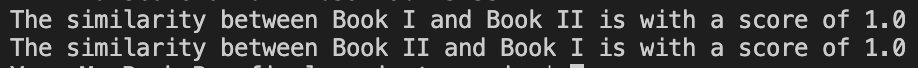
* Book I Chapter XXIV



* Book I Chapter XXV



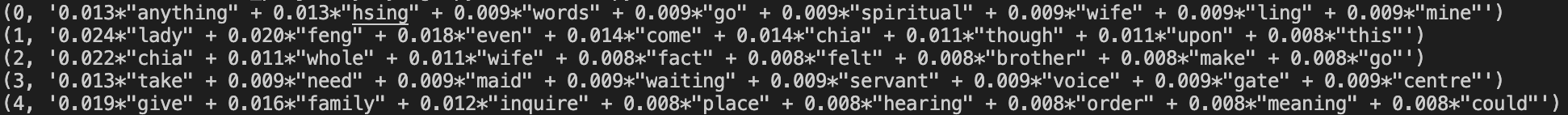
* Overall Book I vs. Book II



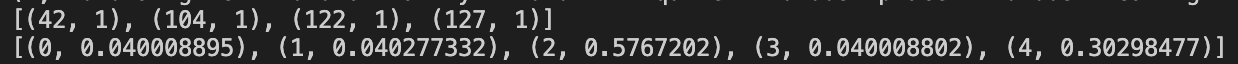
**Theme Analysis**

Try 5 topics in the data

* We print out 8 words from each topic of the data



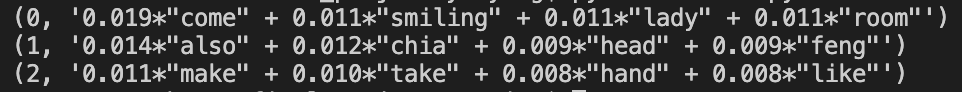
* Topic 2 includes words such as “chia”, “whole”, “wife”, “make”, sounds like a topic related to female dominance. Topic 4 contains words such as “family”, “place”, “order”, sounds like a topic related to family and home.
* Try a new document “The Ladies are the Dominant Figures in the Chia Family, and They Run the House”



* The new document is about female controlling the household. The LDA output shows that topic 2 has the highest probability assigned, and topic 4 has the second highest probability assigned. We agreed!

(remember that the above 5 probabilities add up to 1)

* Connect it to the written text itself [IDENTIFY IT‼‼]
* Find 3 topics in the data using LDA



* The first topic seems to be centered around lady’s rooms, places filled with happiness, harmony, and laughs.
* The second topic seems to suggest the family head figure of Wang Xifeng in the Chia family.
* The third topic hints that the girls in the household are all very intelligent and smart. They not only can use their hand to make useful things, but they can also partake in the intellectual discussions and poetry conversations.

**pyLDAvis Visualization**

pyLDAvis is designed to help users interpret the topics in a topic model that has been fit to a corpus of text data. The package extracts information from a fitted LDA topic model to inform an interactive web-based visualization.

**Saliency**: a measure of how much the term tells you about the topic.

**Relevance**: a weighted average of the probability of the word given the topic and the word given the topic normalized by the probability of the topic.

The size of the bubble measures the importance of the topics, relative to the data.

First, we got the most salient terms, means terms mostly tell us about what’s going on relative to the topics. We can also look at individual topic.

* Visualization with 3 topics

Chart, bubble chart

Description automatically generatedChart, bubble chart

Description automatically generated

Chart, bubble chart

Description automatically generated

* From the size of the bubble, we can see that topic 2 is the most important topic among the three topics listed here, composing 40% of the overall tokens.
* From the term frequency, we can also clearly see the high representation of “lady”, “feng”, “make”, “place” → indication of Lady Feng and Lady Chia controlling the household. It’s interesting that the word “shout” also occurred in a relatively high frequency (might be due to the large household → probably not related to arguments or fights because we also see the representation of “love”, “dear” being frequently mentioned in the book)

Adjusting Relevance for Topic 2:

Relevance is denoted by λ, the weight assigned to the probability of a term in a topic relative to its lift. When λ = 1, the terms are ranked by their probabilities within the topic (the ‘regular’ method) while when λ = 0, the terms are ranked only by their lift. The interface allows to adjust the value of λ between 0 and 1.

Chart, bubble chart

Description automatically generated

* Adjusting the relevance, we see that “make” “lady” “feng” “shout” sticks at the top
* Visualization with 5 topics
* Visualization with 10 topics

Chart, bubble chart

Description automatically generated

* Frequency of each term (in red) and the corpus-wide frequency (in blueish gray). When no topic is selected, the right panel displays the top 30 most salient terms for the dataset.
* For our data, we can see that “lady” “chia” “feng” “make” “people” occurs extremely frequently. This backs up our hypothesis that this large household is run my females – specifically Lady Chia and Wang Xifeng.

Running the same code for Book II, we see the same trend. But we started to see more female figures taking control of the family.

Chart, bubble chart

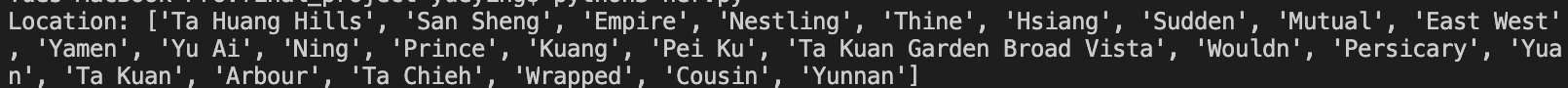
Description automatically generated Chart, bubble chart

Description automatically generated

* We see that Lady Wang, young madame (probably the female servants), and Shi Hsiang-yun occurs more often. From the beginning of the essay, Feng and Lady Chia control the family primarily, to more involvements of the ladies. We see that the ladies are the center and most important figures of the family. IT’S WORTH NOTING that we didn’t not see Pao-yu occurring in any of these graphs!

**NER Analysis** (Named Entity Recognition with SpaCy):

* Location



* The list of locations we get from SpaCy NER analysis is focused on two themes. The royal references/governmental positions (“empire”, “prince”, “Yamen”) and the garden scenes (“hills”, “garden”, “vista”) → seems to also hinting that what happens in the garden has implicit association with the governmental positions and royal power (females lead to the success and prosperity of the family)
* Since many of the ancient Chinese province/city names are very hard to identify using NLP algorithms, sadly, we didn’t identify many city names in the output; thus, it was hard to trace the characters’ movements.

**Semantic Analysis**

****

Likelihood of co-occurrence of two different words

(Pointwise Mutual Information PMI - favors rare association ex. York with New since New York, Log-Likelihood - weights more common associations higher)

Computing top 10 word associations using log-likelihood...

[('pao', 'yu'), ('lady', 'feng'), ('madame', 'wang'), ('hsi', 'jen'), ('tai', 'yu'), ('t', 'un'), ('chia', 'cheng'), ('waiting', 'maid'), ('dowager', 'lady'), ('ch', 'chung')]

[mostly character names]

Computing top 10 word association using PMI...

[('abashed', 'gloomy'), ('abasing', 'natured'), ('abyss', 'afterwards'), ('abyss', 'transmute'), ('accessory', 'predestined'), ('accountant', 'tallied'), ('accurate', 'transcription'), ('accuse', 'wrongly'), ('aconitum', 'ophiopogon'), ('acts', 'annotated')]

[rare words association – output not so useful]

Type a word to get top 10 associated words (CTRL+C to quit): **female**

### Top 10 associated words as measured by Log-likelihood score ###

cousin: 18.686290149502497

male: 6.928918267761864

relative: 4.130061453453928

pupil: 3.9082428886301135

wailing: 3.0484103300866323

earnestly: 2.6664723917156086

contrast: 2.4530095607836726

handiwork: 2.4530095607836726

presuming: 2.4530095607836726

youthful: 2.3041646166825385

### Top 10 associated words as measured by Pointwise Mutual Information ###

wailing: 10.032075681668228

earnestly: 9.032075681668228

contrast: 8.44711318094707

handiwork: 8.44711318094707

presuming: 8.44711318094707

youthful: 8.032075681668228

cake: 8.032075681668228

male: 8.032075681668227

assigned: 7.710147586780865

yore: 7.224720759610623

Type a word to get top 10 associated words (CTRL+C to quit): **power**

### Top 10 associated words as measured by Log-likelihood score ###

expand: 3.1991214973454465

estimate: 3.1991214973454465

kindly: 2.836190719494656

motion: 2.6283833616345915

thorough: 2.6283833616345915

lifetime: 2.2371893337211586

treat: 1.973604235351659

real: 1.9514148691618194

answer: 1.8902808833095408

evening: 1.6405104124801142

### Top 10 associated words as measured by Pointwise Mutual Information ###

expand: 10.55563763772524

estimate: 10.55563763772524

kindly: 9.55563763772524

motion: 8.970675137004083

thorough: 8.970675137004083

lifetime: 7.855197919584149

treat: 7.096206019087943

real: 7.0320756816682275

answer: 6.855197919584148

evening: 6.129372883023143

Type a word to get top 10 associated words (CTRL+C to quit): **woman**

### Top 10 associated words as measured by Log-likelihood score ###

ning: 4.229337201908455

**menage**: 4.075629704274578 (the management of a household)

**matron**: 3.698770199653514 (a woman in charge of domestic and medical arrangements at a boarding school or other establishment)

introduced: 3.178767410853566

maid: 3.1479800506988282

household: 3.0220243559949242

waiting: 2.941830059297645

mansion: 2.563139164422475

argues: 2.535605189970469

memoirs: 2.535605189970469

### Top 10 associated words as measured by Pointwise Mutual Information ###

argues: 8.55563763772524

memoirs: 8.55563763772524

serving: 8.55563763772524

lighted: 8.55563763772524

pre: 8.55563763772524

fascinating: 8.55563763772524

incomparably: 8.55563763772524

conveyed: 8.55563763772524

immodest: 8.55563763772524

indelicate: 8.55563763772524

(Unlike the traditional understanding of women, women in the novel are not often associated with adjectives such as immodest and indelicate. In the entire novel, these two descriptions only occurred once, in very specific settings.)

Type a word to get top 10 associated words (CTRL+C to quit): **women**

### Top 10 associated words as measured by Log-likelihood score ###

pleasing: 3.821144554373697

pure: 3.006531762211936

eye: 2.0814911536092047

say: 1.9478460213448878

### Top 10 associated words as measured by Pointwise Mutual Information ###

pleasing: 12.233709542837879

pure: 9.911781447950517

eye: 7.242754682440885

say: 6.855197919584148

Type a word to get top 10 associated words (CTRL+C to quit): **lady**

### Top 10 associated words as measured by Log-likelihood score ###

feng: 509.6832008295258

chia: 77.2489985601394

secunda: 13.580001074810454

wang: 11.883763596608024

madame: 10.377772468911601

apartment: 5.972435878156754

sent: 5.29143946839517

inquired: 4.440318253499711

laughed: 4.042221195997346

repast: 3.9573327339038817

### Top 10 associated words as measured by Pointwise Mutual Information ###

slackened: 4.888526095650214

tractable: 4.888526095650214

intercede: 4.888526095650214

perfection: 4.888526095650214

plump: 4.888526095650214

associated: 4.888526095650214

spoilt: 4.888526095650214

scurry: 4.888526095650214

usually: 4.888526095650214

warned: 4.888526095650214

Type a word to get top 10 associated words (CTRL+C to quit): **family**

### Top 10 associated words as measured by Log-likelihood score ###

estate: 10.462692612364803

school: 9.667720991804188

companion: 8.56638801441462

official: 6.5331293893272075

generation: 5.067095490168622

connection: 5.014244723568944

banquet: 4.967787639192766

tie: 3.91861003570778

connexion: 3.6631341471839596

circle: 3.6631341471839596

### Top 10 associated words as measured by Pointwise Mutual Information ###

cultured: 6.5167186484329385

actual: 6.5167186484329385

sustain: 6.5167186484329385

chattel: 6.5167186484329385

uninterrupted: 6.5167186484329385

reckoning: 6.5167186484329385

flourishing: 6.5167186484329385

salacious: 6.5167186484329385

perfection: 6.5167186484329385

disgrace: 6.5167186484329385

**Word Embedding & Word2Vec**

Word Embedding: Collective term for models that learned to map a set of **words** or phrases in a vocabulary to **vectors** of numerical values

Word2Vec: group the vectors of similar words together in vectorspace. That is, it detects similarities mathematically. Word2Vec is not a singular algorithm, rather, it is a family of model architectures and optimizations that can be used to learn **word embeddings** from large datasets.

Chinese Version of the Text: [spaCy](https://spacy.io/models/zh#zh_core_web_trf)