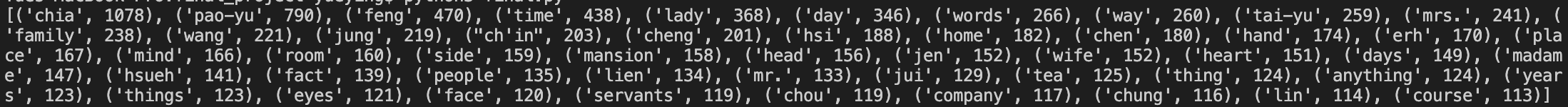
DRC Textual Analysis Output

* 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



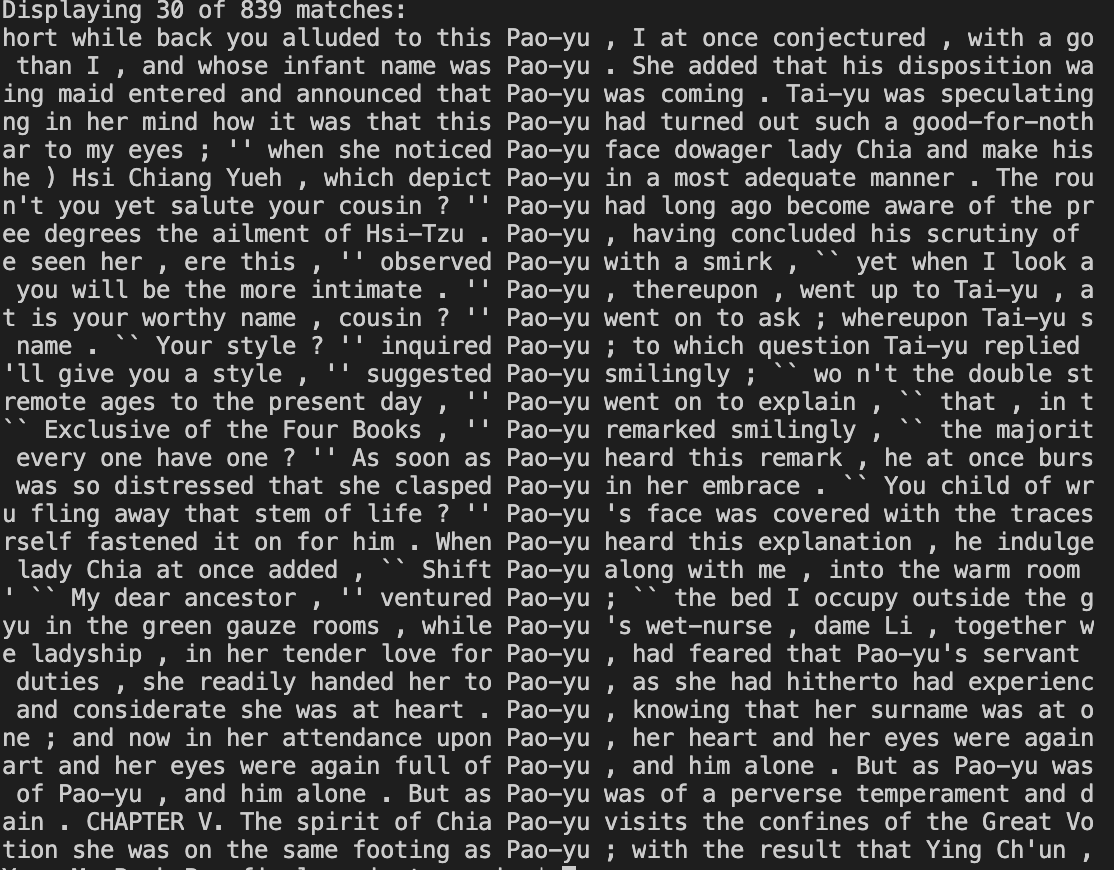
* We see that when薛is mentioned, it is always mentioned in a similar setting as林and贾combined. It seems that she is naturally the third-wheeler in林黛玉and贾宝玉’s relationship. “they” occurs very frequently in a similar text as薛, indicating that薛is mentioned as a singular unit while林and贾are mentioned as a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



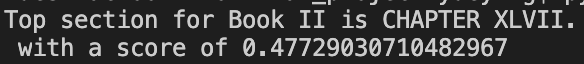
* Book I Chapter II



* Book I Chapter III



* Book I Chapter IV



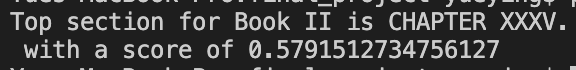
* Book I Chapter V



* Book I Chapter VI



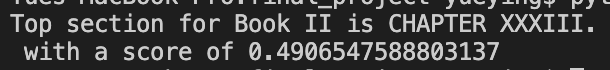
* Book I Chapter VII



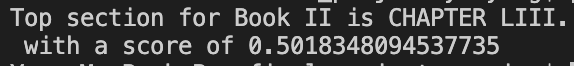
* Book I Chapter VIII



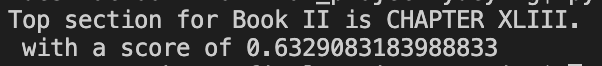
* Book I Chapter IX



* Book I Chapter X



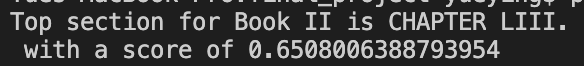
* Book I Chapter XI



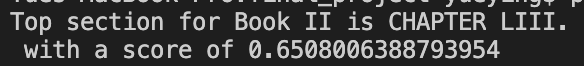
* Book I Chapter XII



* Book I Chapter XIII



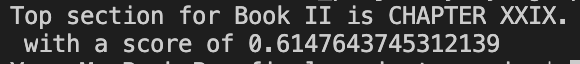
* Book I Chapter XIV



* Book I Chapter XV



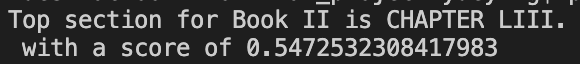
* Book I Chapter XVI



* Book I Chapter XVII



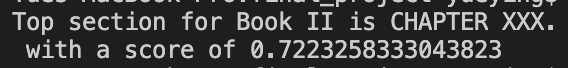
* Book I Chapter XVIII



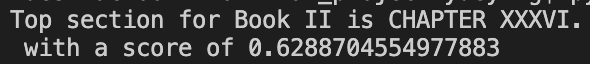
* Book I Chapter XIX



* Book I Chapter XX



* Book I Chapter XXI



* Book I Chapter XXII



* Book I Chapter XXIII



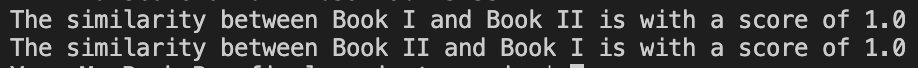
* Book I Chapter XXIV



* Book I Chapter XXV



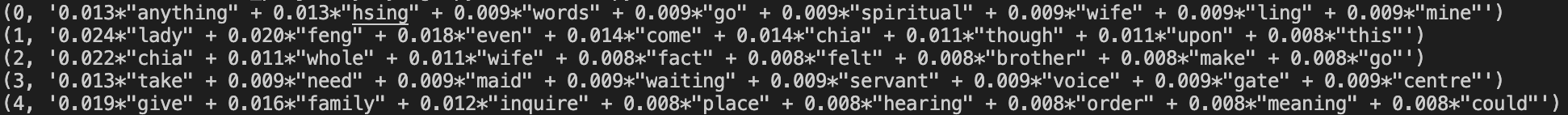
* Overall Book I vs. Book II



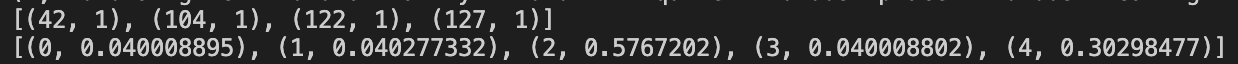
**Theme Analysis**

Try 5 topics in the data

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



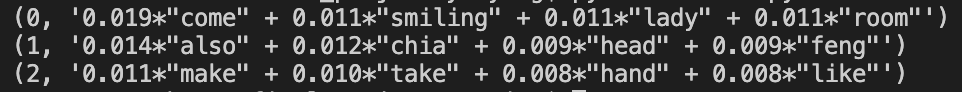
* Topic 2 includes words such as “chia”, “whole”, “wife”, “make”, sounds like a topic related to female dominance. Topic 4 contains words such as “family”, “place”, “order”, sounds like a topic related to family and home.
* Try a new document “The Ladies are the Dominant Figures in the Chia Family, and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

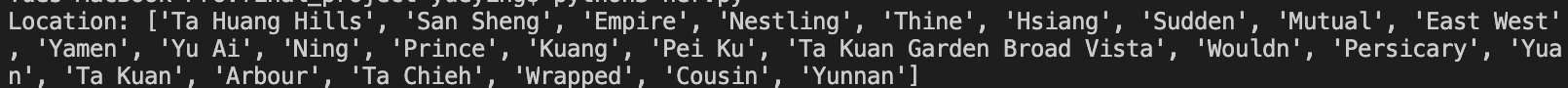
Description automatically generated Chart, bubble chart

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 (…The bones and flesh of woman, he argues, are made of water, while those of man of mud. 'Women to my eyes are **pure and pleasing**,' he says, 'while at the sight of man, I readily feel how **corrupt, foul and repelling** they are! ...)

from Paoyu – strongly advocates for women

memoirs: 8.55563763772524 ("Memoirs of **excellent** women,")

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. Instead, they are associated with excellence, pure and pleasing. Men are being despised.)

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.

**Close Reading Textual Evidence**

贾宝玉女性人格特征：

“贾宝玉是具有初步男人解放理念的人物形象，体现了曹雪芹思想中的女性主义与男性解放主义意识。**在贾宝玉身上我们不但可以看到他秉具男人的人格特征，同时也可以看到秉具女性的人格特征**。”[6]他对女性的尊崇和对男人的贬抑，使他自觉或不自觉地以女性的品性标准来规范自己，进而使他的言行表现出较明显的女性化特征。贾宝玉是一个女权至上论者，他同情女性、崇拜女性、讴歌女性，而对于男子，则充满了鄙夷、厌恶和痛恨，并为自己身为男子而耻。他认为“女儿是水做的骨肉，男人是泥做的骨肉，我见了女儿便清爽，见了男子便觉浊臭逼人”（第二回）。他与女性交往的平等观念，以及对女性的爱护，通过许多细节表现出来。

**女性的不同凡响：**

对作者而言，女性的理想人格意味着把一种引导精神出路的探索合二为一。宝玉认为：“强似前代书中所有之人”的，并非齐家治国的伟丈夫，而是几个异样女子。她们看似平常，“或情或痴”，“或小才微善”，实则不同凡响。因为她们不属于男权文化的产品，“亦天班姑，蔡女之德能”。宝玉为女性伸张人格，张扬女性的价值权利。

Text：第一回

空空道人遂向石头说道: "石兄,你这一段故事,据你自己说有些趣味,故编写在 此,意欲问世传奇.据我看来,第一件,无朝代年纪可考,第二件,**并无大贤大忠理朝廷 治风俗的善政**, 其中**只不过几个异样女子,或情或痴,或小才微善**,亦无班姑,蔡女之德能.我纵抄去,恐世人不爱看呢."石头笑答道:"我师何太痴耶!若云无朝代可考,今我师竟假借汉唐等年纪添缀, 又有何难?但我想,历来野史,皆蹈一辙,莫如我这不借此套者, 反倒新奇别致,不过只取其事体情理罢了,又何必拘拘于朝代年纪哉!再者,市井俗人喜看理治之书者甚少,爱适趣闲文者特多.历来野史,或讪谤君相,或贬人妻女,奸淫凶恶,不可胜数.更有一种风月笔墨,其淫秽污臭,屠毒笔墨,坏人子弟,又不可胜数.至若佳人才子等书,则又千部共出一套,且其中终不能不涉于淫滥,以致满纸潘安,子建,西子,文君,不过作者要写出自己的那两首情诗艳赋来,故假拟出男女二人名姓,又必旁出一小人其间拨乱,亦如剧中之小丑然.**且鬟婢开口即者也之乎,非文即理**.故逐一看去,悉皆自相矛盾,大不近情理之话,竟不如我半世亲睹亲闻的这几个女子,虽不敢说**强似前代书中所有之人**,但事迹原委,亦可以消愁破闷,也有几首歪诗熟话,可以喷饭供酒.

**女性掌权**

**第十五回**：王凤姐弄权铁槛寺 秦鲸卿得趣馒头庵

宝玉又将北静王所赠鹡鸰香串珍重取出来，转赠黛玉。黛玉说： “什么臭男人拿过的！我不要他。”遂掷而不取。宝玉只得收回，暂且无话

**第十五回**：王凤姐弄权铁槛寺 秦鲸卿得趣馒头庵

凤姐也便回至净室歇息，老尼相伴。此时众婆子媳妇见无事，都陆续散了自去

歇息，跟前不过几个心腹小丫头，老尼便趁机说道：“我有一事，要到府里求太太，

**先请奶奶的示下**。”凤姐问道：“什么事？”老尼道：“阿弥陀佛!只因当日我先

在长安县善才庵里出家的时候儿，有个施主姓张，是大财主。他的女孩儿小名金哥，

那年都往我庙里来进香，不想遇见长安府太爷的小舅子李少爷。那李少爷一眼看见

金哥就爱上了，立刻打发人来求亲，不想金哥已受了原任长安守备公子的聘定。张

家欲待退亲，又怕守备不依，因此说已有了人家了。谁知李少爷一定要娶，张家正

在没法，两处为难；不料守备家听见此信，也不问青红皂白，就来吵闹，说：‘一

个女孩儿你许几家子人家儿？’偏不许退定礼，**就打起官司来**。女家急了，只得着

人上京找门路，赌气偏要退定礼。**我想如今长安节度云老爷，和府上相好，怎么求**

**太太和老爷说说，写一封书子，求云老爷和那守备说一声，不怕他不依**。要是肯行，

张家那怕倾家孝顺，也是情愿的。”凤姐听了笑道：“这事倒不大。**只是太太再不**

**管这些事**。”老尼道：“**太太不管，奶奶可以主张了**。”

* 政府的事，衙门律法上也都是贾府说了算，并且不是贾府的男人，是贾府的女人（先前是老太太-贾母，现在是奶奶-王熙凤）
* 女人的权利不仅仅是在家庭内部

**第十三回**： 秦可卿死封龙禁尉　王熙凤协理宁国府

…贾珍笑道：“婶娘的意思侄儿猜着了，是怕大妹妹劳苦了。若说料理不开，从小儿大妹妹玩笑时就有杀伐决断，如今出了阁，**在那府里办事，越发历练老成了**。我想了这几日，**除了大妹妹再无人可求了**。婶娘不看侄儿和侄儿媳妇面上，只看死的分上罢！”说着**流下泪**来。

…贾珍见凤姐允了，又陪笑道：“也管不得许多了，**横竖要求大妹妹辛苦辛苦**。我这里**先与大妹妹行礼**，等完了事，我再**到那府里去谢**。”说着就作揖，凤姐连忙还礼不迭。

　　贾珍便命人取了宁国府的对牌来，命宝玉送与凤姐，说道：“**妹妹爱怎么就怎么样办，要什么，只管拿这个取去，也不必问我**。只求别存心替我省钱，要好看为上；二则也同那府里一样待人才好，不要存心怕人抱怨。只这两件外，我再没不放心的了。”

* 贾珍的妻子秦可卿去世了，他却求着王熙凤帮他料理后事。王熙凤是荣国府的人，并非宁国府的人，但是宁国府宁可让她来料理事情，王熙凤不仅掌权荣国府还掌权宁国府，威严决断

除此以外，贾母是贾府的祖宗，地位至高无上，太多太多的小例子。

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