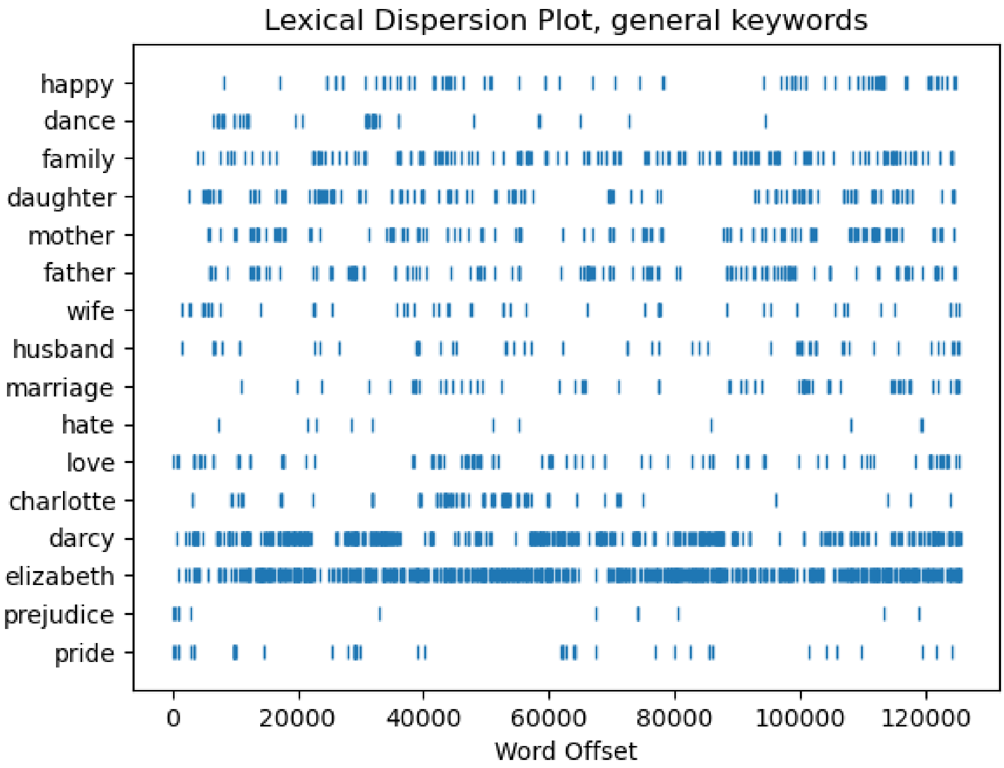
Excercise set 3 DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercise 2.5 (Gutenberg_crawler.ipynb)

Code is in file Gutenberg_crawler.ipynb, here's top 100 words after pruning:

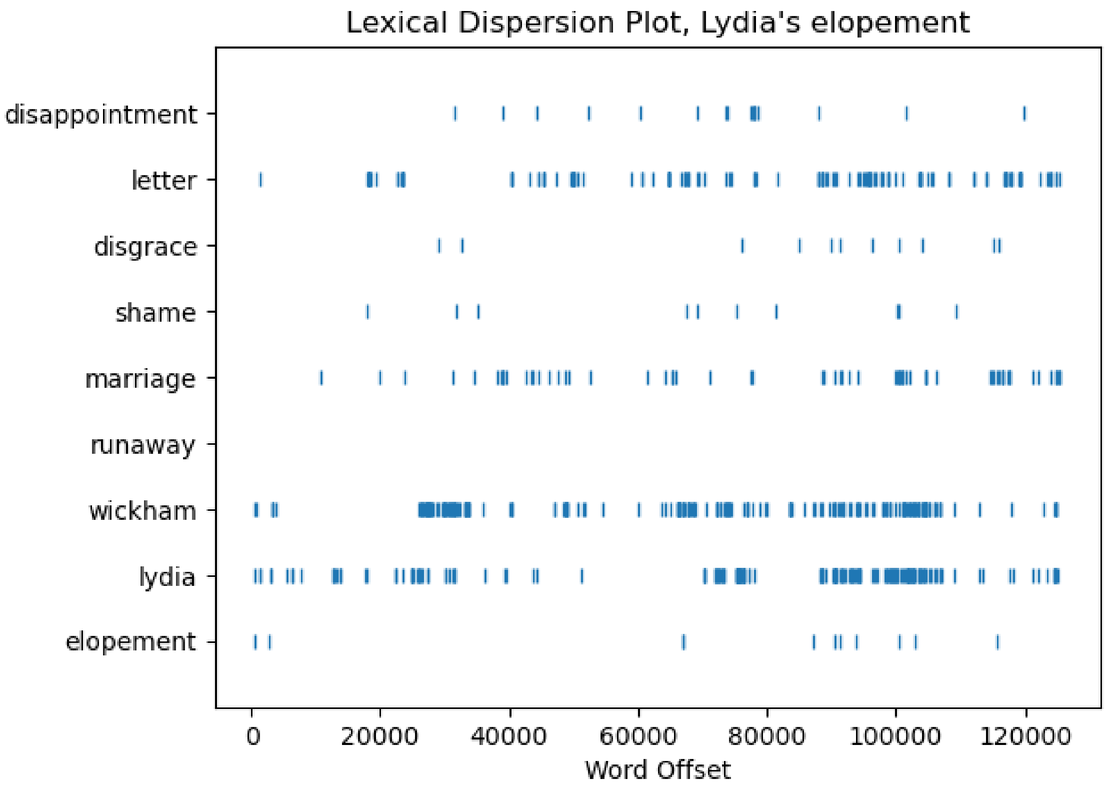
```
Pruned top 100 words in the downloaded
books:
passed: 711
possible: 709
service: 709
circumstance: 708
four: 706
visit: 705
fine: 703
please: 701
age: 699
second: 698
peace: 698
Meg: 698
news: 696
wrong: 695
passion: 695
mistress: 694
met: 690
England: 688
Have: 687
madam: 687
fell: 685
conversation: 682
One: 681
sound: 680
Monsieur: 680
table: 679
earth: 678
seeing: 678
spoke: 677
sleep: 677
hast: 676
seem: 674
interest: 674
fool: 674
Aramis: 674
white: 673
view: 668
smile: 666
behind: 665
appeared: 664
creature: 663
cousin: 661
youth: 660
From: 660
became: 660
enemy: 660
Allworthy: 660
taking: 656
affair: 652
object: 652
seems: 651
Amy: 651
several: 648
Paris: 648
certainly: 646
London: 646
story: 646
music: 644
captain: 644
hundred: 642
ground: 642
large: 638
strong: 637
duty: 637
Our: 636
cold: 636
affection: 636
afraid: 632
run: 631
act: 630
glad: 629
bad: 629
pay: 629
Casaubon: 629
suppose: 626
window: 625
along: 624
virtue: 622
John: 619
Nor: 618
ill: 618
sun: 614
married: 614
entered: 614
piece: 613
neither: 612
ought: 610
Shall: 609
France: 608
devil: 606
Laurie: 604
hair: 603
gentle: 603
lip: 601
regard: 601
war: 600
Fred: 599
sometimes: 598
favour: 596
added: 595
```

Exercise 3.1 (pride_and_prejudice.ipynb)

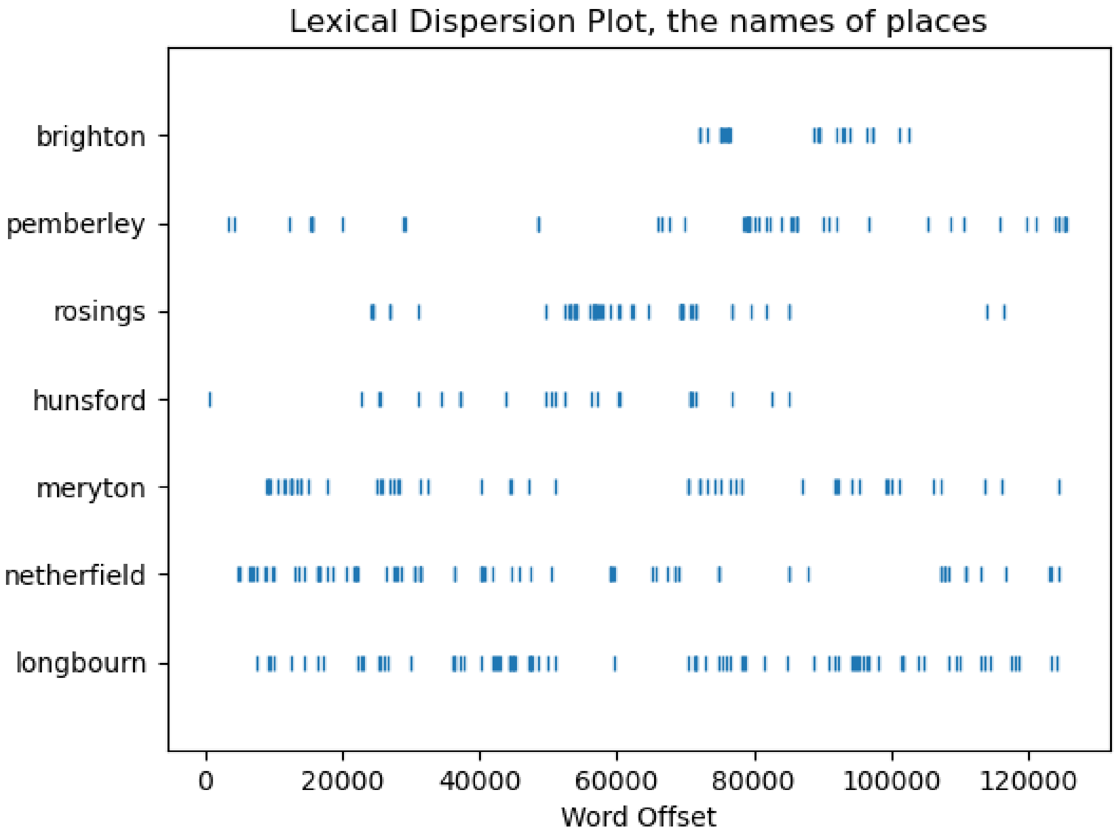


The book has a happy ending, as Elizabeth and her oldest sister, Jane, get married right at the end of the story. This explains why the word "happy" appears more frequently in the latter part of the book. Words like "mother," "father," "family," and "daughter" are present throughout the entire narrative since family is a central theme; the main character, Elizabeth, has four sisters, as well as her mother and father.

There are two significant dances at the beginning of the story: one where the Bennet family meets Mr. Darcy and Mr. Bingley, and another where the Bennet family humiliates themselves. Charlotte’s name appears more often in the beginning and the middle of the book, which is likely when Elizabeth goes to visit her and Mr. Collins.



I also want to create a lexical dispersion plot focused on Lydia Bennet's elopement. Words like "elopement," "marriage," "Lydia," "Wickham," "disgrace," and "letter" occur more frequently in this section of the book, which is around 90,000 words.



The names of places also tell a lot about the story. Brighton is mentioned for the first time when Lydia expresses her desire to go there because of the military presence, and it is referenced again when she elopes with Mr. Wickham. Darcy's mansion, Pemberley, is introduced early on when his wealth is discussed, but it becomes even more significant when Elizabeth visits it. Rosings is mentioned at the beginning when Mr. Collins brags about how wealthy Lady Catherine de Bourgh is, and it is discussed further when Elizabeth visits. Hunsford is mentioned in the same context, as Elizabeth is there when she goes to Rosings Park. Meryton, being a nearby town, is frequently mentioned, although it is less prominent when Elizabeth is visiting Charlotte and Mr. Collins. Netherfield Park is Mr. Bingley's mansion, while Longbourn is the Bennet family's home, so both are referenced throughout the book.

I know the plot quite well without asking ChatGPT so I'm skipping the part d of this exercise.

Exercise 3.2 (Frankenstein.ipynb)

"Science" appears in sentences connected to Victor's obsession with scientific discovery and his ambition to become a remarkable scientist.

"Horror" reflects Victor's emotional response to his creation; capturing the terror and regret he feels after bringing the monster to life.

"Monster" illustrates the intense disgust Victor harbours toward the being he has created, emphasizing his inner turmoil.

"Fear" often follows "my" or "I", indicating how deeply personal Victor's fear is. He dreads both the monster and the devastating consequences of his own actions.

I asked ChatGPT "How is science featured in Frankenstein?"

ChatGPT's reply consisted of:

- Ambition and the pursuit of knowledge
- the creation of life
- isolation and alienation
- consequences of scientific exploration
- critique of scientific endeavours
- personal fear

Both my analysis and ChatGPT's reply emphasize Victor's relentless ambition and pursuit of knowledge. His desire to be a remarkable scientist drives the narrative and is a central aspect of his character development. I discuss in my analysis how Victor feels terror after bringing monster to life. This reflects how the theme of creation intertwines with Victor's emotional turmoil after he realizes the consequences of his scientific endeavours. Isolation and alienation weren't really mentioned in my analysis. Consequences of scientific exploration is mentioned in my analysis when I'm discussing how deep Victor's fear is.

Exercise 3.5

document feature category	my addition
Length of the document	Number of headings and subheadings
Metadata	Keywords/Tags
Connectivity	Number of Social Media Shares

Popularity	Number of Comments
Sentiment	Subjectivity
Reception	Upvote/downvote ratio