Practice Assignment #2
Text Analysis & Visualization

General Instructions

Use the Austen novels dataset (available in Voyant Tools and through the attached file) to create visualizations and answer the questions below. Please include your visualizations and narrative in this assignment document. Each of your visualizations will be evaluated based on the accuracy of information display, thoughtfulness of visual design, and inclusion of necessary graph components (e.g., title, labels, legends).

Please submit your assignment to Canvas. If you create the visualizations with Voyant Tools, you **only** need to submit the assignment document with your visualizations and narrative, and do not need to include source files. For other tools, please include the original visualization file, in addition to the assignment document.

Question #1

Create a word cloud that displays the top 90-100 keywords from *Pride and Prejudice* in the Austen novels dataset. You may decide if you want to remove any stop words (or not).
After you complete the visualization, use a couple of sentences to describe the word cloud, and list the top 5 keywords with their frequencies.

Ans:

To begin this task, I first created a separate **Cirrus visualization** specifically for *Pride and Prejudice* from the Austen novels dataset in Voyant Tools. I adjusted the **terms setting to 100**, as required by the assignment, to ensure the word cloud included a dense and representative sample of the most important keywords.

I also made a few formatting adjustments to improve the overall appearance and clarity. I selected **Auto-detect** for stop words to filter out common English words that add little meaning (e.g., "the," "is"), allowing content-specific terms to stand out. I changed the **font to Georgia** to give the cloud a more literary aesthetic that better matches the tone of the novel, and used a **circular layout** for a more polished and balanced look.

The word cloud highlights dominant terms such as **"elizabeth," "mr," "darcy," "mrs,"** and **"bennet"**, which reflect the importance of these characters in the narrative. Dialogue words like **"said"** appear prominently, capturing Austen's conversational writing style. Other keywords such as **"know," "think," "family,"** and **"sister"** reinforce major themes of self-awareness, social expectations, and interpersonal relationships.

Below, Figure 1.1 displays the top 100 keywords using a customized Cirrus panel in Voyant Tools.

Vaishnavi Pawar
vpawar@iu.edu

Figure 1.1: Word Cloud - Top 100 Keywords in Pride and Prejudice

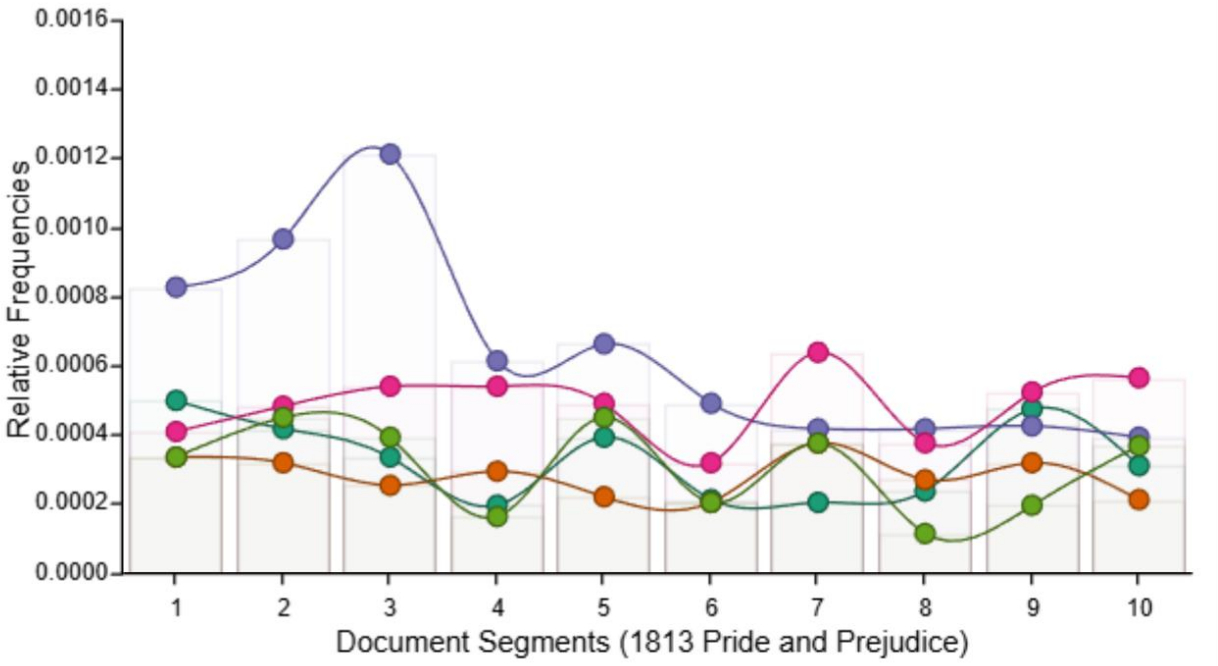
I included Figure 1.2 to support my word cloud by showing the **actual word counts** of the most frequently used words. These frequencies confirm that the words highlighted in the visualization are truly the most common in the novel. This also **directly addresses the assignment's requirement** to list the top five keywords with their frequencies, using concrete data to back up the visual display.

Figure 1.2: Top 5 Keywords with Frequencies

Cirrus | Terms | Links

	Term	Count
1	mr	785
2	elizabeth	597
3	said	401
4	darcy	373
5	mrs	343

Figure 1.3: Word Frequency Trends Across the Novel



To further explore word usage over time, I generated a trend line showing how frequently the top terms appeared across **ten equal-length segments** of the novel. I observed that **"mr"** peaks early in the book, likely corresponding with the formal introduction of key characters - while **"elizabeth"** remains consistently high throughout, confirming her presence as the central protagonist. **"Said"** shows a steady pattern, reinforcing the book's dialogue-driven structure. This visualization supports my **word cloud** as well as my **top five keyword analysis** - by showing not only how often these words appear overall, but also how their usage evolves throughout the novel.

Question #2

(1) Create a graph to show the distribution of the two top verbs, "said" and "think" across different Austen novels.

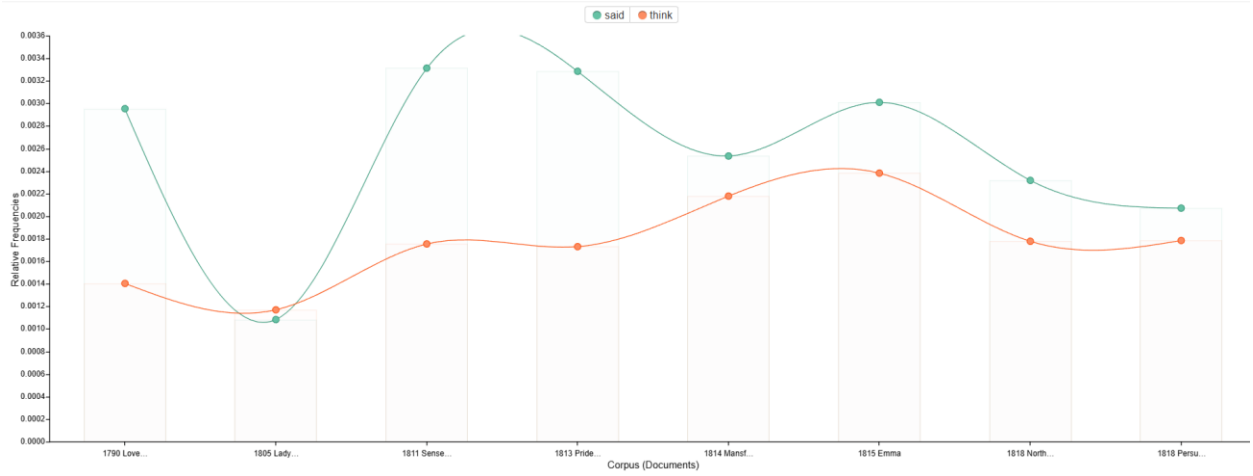
After you complete the visualization, use a couple of sentences to discuss the rationale for your visualization, and answer the following questions:
(1) Which novel(s) has the highest proportion of using "said" in the text?
(2) Which novel(s) has the highest proportion of using "think" in the text?

Ans:

To complete this task, I switched back to the full Austen corpus instead of just Pride and Prejudice. I then used the *Trends* tool in Voyant Tools to visualize the distribution of the two top verbs, **"said"** and **"think,"** across all eight novels. I applied a filter to include only these two terms in the chart, so the visualization could focus *exclusively* on their specific usage patterns without additional clutter.

To improve the clarity and effectiveness of the visualization, I adjusted the settings to display **relative frequencies**, which represent the **proportion of each word relative to the total word count** in each novel. I also changed the color palette to **Set2** for better contrast and kept **stopwords set to Auto-detect** to filter out common filler words. Finally, I selected the **Line + Stacked Bar** chart type to make the differences in word proportions more visually apparent.

Fig 2.1 Distribution of the two top verbs, "said" and "think" across different Austen novels.



The chart (Figure 2.1) clearly shows that the **highest proportion of "said"** occurs in **Sense and Sensibility (1811)** with a relative frequency of **0.00331**, closely followed by **Pride and Prejudice (1813)** at **0.00328**. This indicates that these novels are particularly rich in dialogue. In contrast, the **highest proportion of "think"** appears in **Emma (1815)** at **0.00238**, which may reflect the novel's emphasis on introspection and inner thought.

By focusing on proportional use rather than raw counts, this visualization allows for a fair comparison between texts of different lengths. It directly answers the assignment's questions by using clear evidence to identify which novels rely most heavily on each of these verbs.

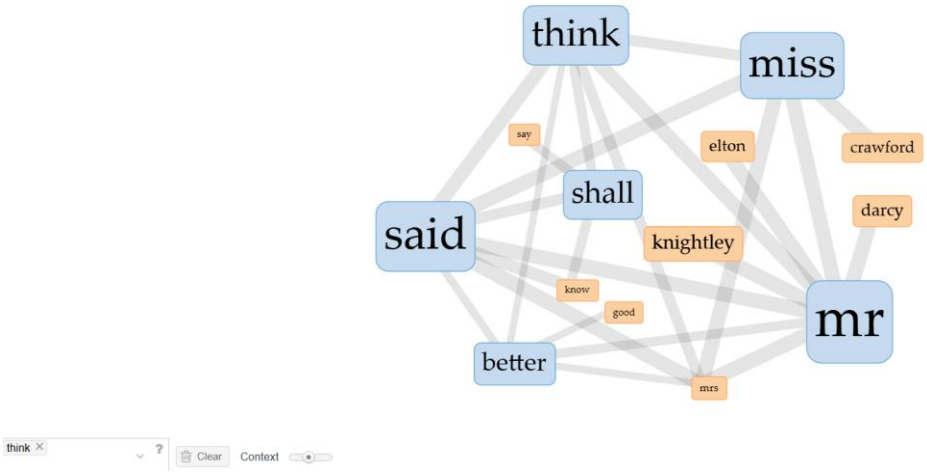
(2) Visualize the collocation network of "think" in the Austen novels, at the context level of 15.

After you complete the visualization, use a couple of sentences to discuss the rationale for your visualization and the information you observed from the collocation graph.

Ans:

To complete this task, I used the *Collocates* tool in Voyant Tools and filtered it specifically for the word **"think."** I also set the context window to **15 words**, as instructed, to capture a broader linguistic environment around the target word. I kept the stopwords option on **Auto-detect** so that only semantically meaningful words appeared in the network. This ensured that common, less informative words like "the" or "and" wouldn't dominate the visualization.

Figure 2.2 Collocation network of "think" in the Austen novels, at the context level of 15.



The collocation graph (Figure 2.2) reveals that **"think"** appears most frequently alongside words like **"shall," "said," "miss," "mr," "better,"** and names such as **"knightley," "elton,"** and **"darcy."** These collocates suggest a pattern of introspection and social dialogue, as characters are shown thinking about decisions, conversations, and people around them. The appearance of honorifics like "mr" and "miss" further supports the idea that thoughts are often directed toward others, reflecting social norms and romantic tension in Austen's novels.

This visualization helped me understand not just how often "think" is used, but how it interacts with other language in Austen's writing. It adds depth to the previous analysis by showing the types of words and characters that co-occur with internal reflection in the text.

Question #3

Use the topic modelling (LDA) and visualization tool in Voyant to explore Austen's novels. Create visualizations to address the following questions:

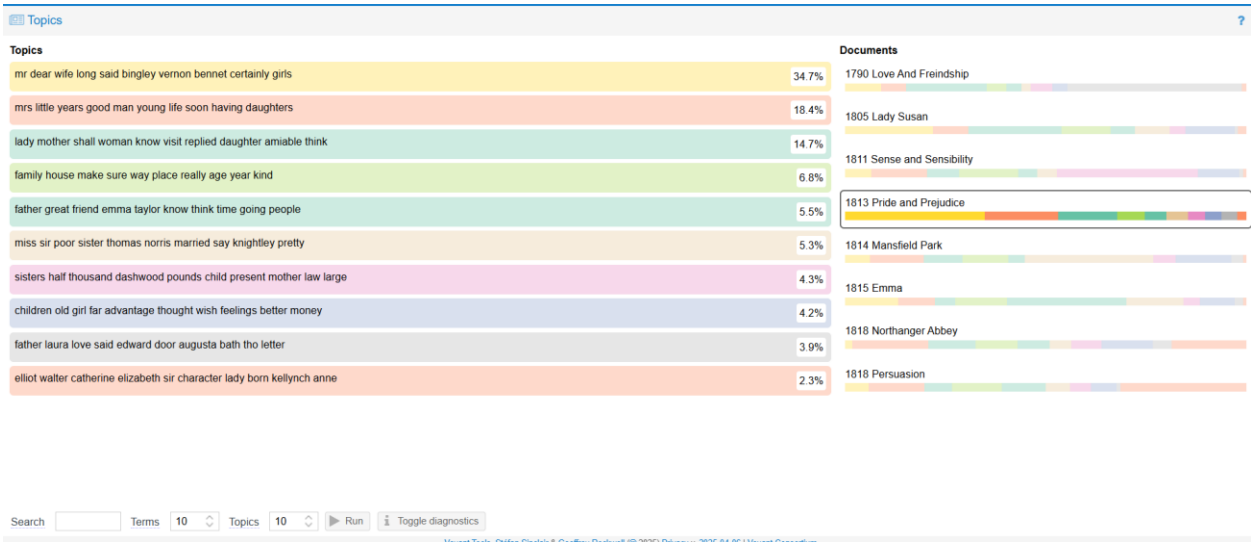
(1) What are the top 10 topics in *Pride and Prejudice*? Please include top 5 representation words for each topic.

Ans:

To complete this task, I used the **Topics panel in Voyant Tools** with **LDA topic modeling** enabled. I filtered the view to **Pride and Prejudice** from the full Austen corpus and set the number of topics to **10** and **terms per topic to 10** as required. Each colored bar in the interface represents the proportional presence of a topic in a particular document. I hovered over the bar corresponding to *Pride and Prejudice* to identify which topics were most prominent in the novel.

Vaishnavi Pawar
vpawar@iu.edu

Figure 3.1: Top 10 topics in *Pride and Prejudice*



I then chose the **top 5 representative words** for each topic, as required. This allowed me to focus on the most prominent conceptual themes within the novel.

By analyzing the **color-coded proportions** in the topic modeling output, I was able to identify which LDA topics were most relevant to *Pride and Prejudice*. The most dominant themes included words like **"mr," "said," "wife,"** and **"bennet,"** which align well with the novel's dialogue-heavy structure and its focus on characters and social relationships.

The following table lists the **top 10 topics** in *Pride and Prejudice*, along with their **top 5 representation words**, based on the colored segments of the LDA output:

Topic #	Top 5 Words in Topic
1	mr, dear, wife, long, said
2	mrs, little, years, good, man
3	lady, mother, shall, woman, know
4	family, house, make, sure, way
5	father, great, friend, emma, taylor
6	miss, sir, poor, sister, thomas
7	sisters, half, thousand, dashwood, pounds
8	children, old, girl, far, advantage
9	father, laura, love, said, edward
10	elliot, walter, catherine, elizabeth, sir

These topics reflect the novel's key themes such as **family structures**, **gender roles**, **social class**, **romantic relationships**, and **economics**.

It's also important to clarify that these words are **not listed by raw frequency**. Instead, Voyant uses **Latent Dirichlet Allocation (LDA)** to identify how strongly a word is associated with a given topic using the **term-topic probability distribution (β)**. The words at the beginning of each list are those most strongly tied to that topic, offering a deeper insight into what each theme represents. **Hence, I used the top 5 representation words for each topic, as the assignment instructed.**

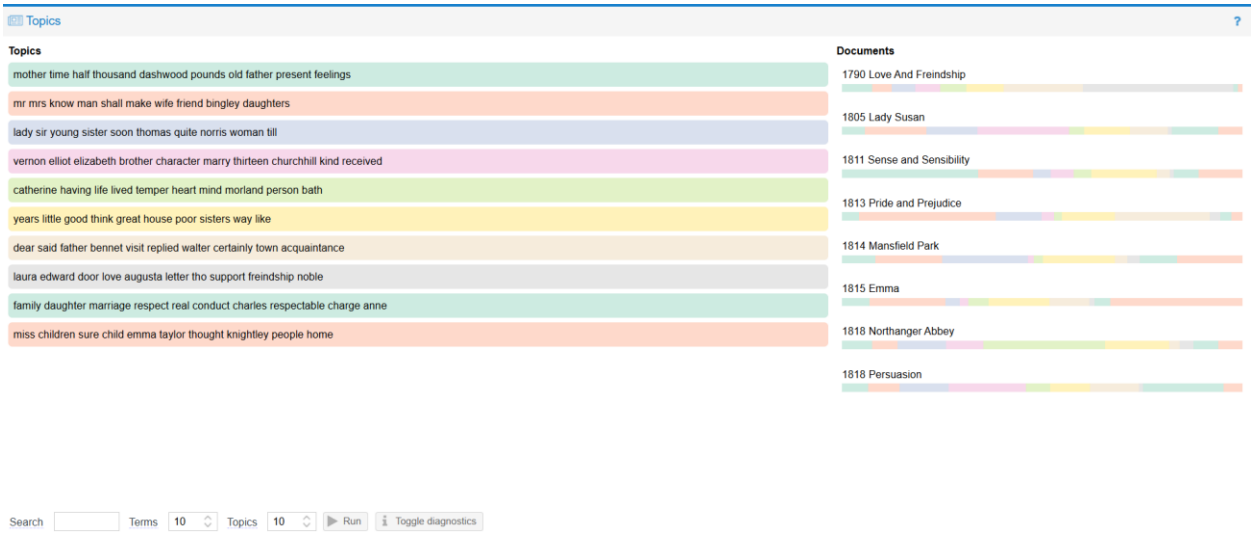
(2) Use the topic modeling results to analyze the differences and similarities between the eight novels by Jane Austen in the corpus.

Ans:

To analyze the similarities and differences among all eight novels in the Austen corpus, I used Voyant Tools' topic modeling feature with LDA enabled. I set the number of topics to 10 and the number of terms per topic to 10. I did not filter by a specific novel, so the visualization reflects patterns across the full corpus.

Figure 3.2 below shows each of the 10 LDA-generated topics (left) and how prominently each topic appears in each novel (right) as a horizontal color-coded bar. Each bar segment's length indicates the proportion of the document that is associated with a particular topic.

Figure 3.2 10 LDA-generated topics of Austen's novels



Similarities Across Novels:

- Certain themes appear consistently across many novels. For example, the **green topic** containing words like *"mother," "time," "half," "father," "feelings"* is present in nearly all books, indicating a shared thematic focus on family dynamics and emotions.

- The **peach-colored topic** with words like *"mr," "mrs," "know," "wife," "friend," "bingley," "daughters"* is another recurring theme. This suggests that social relationships and interactions are a common thread throughout Austen's work.
- The **yellow topic** including *"years," "good," "think," "great," "house," "sisters," "way," "like"* also spans several novels and seems to capture everyday domestic life and moral judgments, another signature of Austen's style.

Differences Between Novels:

Some topics appear more dominant in specific novels, highlighting the differences in themes and focus across the Austen corpus. For example:

- The **orange topic**, which includes terms like *"emma," "taylor," "thought," "knightley," "people"*, appears most prominently in **Emma (1815)**. These terms clearly reflect the novel's focus on introspection and character-driven social interaction.
- The **purple topic**, featuring *"vernon," "elliot," "elizabeth," "character," "churchhill," "received"*, is strongly represented in both **Persuasion (1818)** and **Lady Susan (1805)**. This suggests a shared emphasis on named characters, complex relationships, and social correspondence.
- The **grey topic**, with words like *"laura," "edward," "augusta," "letter," "support," "noble"*, is dominant in **Love and Friendship (1790)**. These words point to the novel's epistolary style and focus on romantic and familial duty.

By comparing the color-coded topic proportions across the eight novels, I was able to identify both unique and recurring themes in Austen's works. Some topics - like family dynamics and social titles - are widely distributed across multiple novels, showing thematic overlap. Others are more distinct and help define the central focus of individual works.