Title: Analysis of Thematic Differences between 19th-Century Slave Narratives and Romance Fiction

1. My Dataset

I used texts from Project Gutenberg, comparing two different categories of novels: the slave theme and the romance theme. The issue is that the books I chose each have only about 30–40 chapters to make it 100 chapters at least for each category, so I had to combine three different books per category, which means I ended up using six books in total. I'm hoping to see clear differences between the two distinct genres.

Category Slave: The Garies and Their Friends, 1857, 36 chapters, Uncle Tom's Cabin, 1852, 43 chapters, Clotelle, 1867, 28 chapters

Category Romance: Emma, 1815, 55 chapters, The Age of Innocence, 1920, 34 Chapters, Wuthering Heights, 1847, 34 Chapters

I know that the chapters in each book aren't the same length, but for now it's easier to use chapters as the unit. My dataset ended up with 243 documents, rather than the original 230 chapters. I checked that none of the documents contains fewer than 500 words, just in case the text was split at an awkward point. So I think the dataset is good enough to proceed with the analysis. There are 120 files in Slave_Fiction, with an average of 2,953 words per file, and 123 files in Romance_Fiction, with an average of 3,045 words per file.

2. Processing Data

First, I removed the Gutenberg headers and footers from the texts downloaded from Project Gutenberg. Then, I performed tokenization and converted all words to lowercase. I also applied a stemmer.

This time, I used a simple split() function for tokenization instead of word_tokenize, which I used in the previous assignment. Also did not use isalpha(). The text does not look as clean at this stage, and I am not sure whether this is due to the characteristics of the novels or the tools I used.

I manually filtered the words by keeping only those with a length greater than two characters. I believe we do not need conversational sounds such as "ah," "um," and "oh" in the analysis. However, due to issues in my code—possibly hidden punctuation, extra spaces, or other invisible characters—some unwanted tokens still entered the dataset. As a result, I noticed words like "ma" appearing in the final output. Next, I calculated the Log-Likelihood Ratio (LLR) using the formula from the assignment. I suspected that character names might be hiding the real themes, so I created a new version of the data without names to see how the counting results changed. After that, I moved on to Latent Dirichlet Allocation (LDA). I experimented with removing different groups of words to make the topics clearer. Finally, even though I wasn't sure if TF-IDF was necessary, I decided to test it. This turned out to be very useful because it allowed me to compare how different models handle important words.

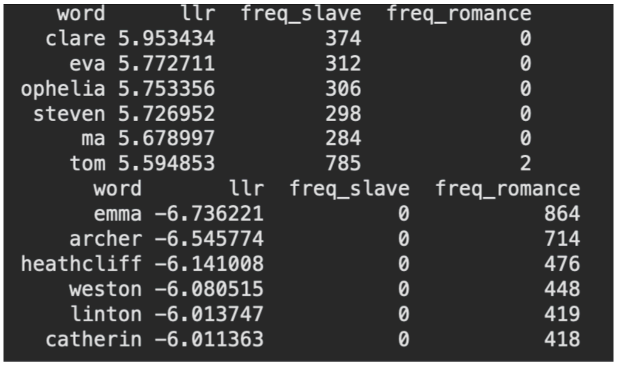
For experimented with removing different groups of words. Initially, I used standard English stop words, but character names like "Eva" and "Emma" dominated the topics. I used an iterative approach: I ran the model, checked the top words, and manually added those names and common verbs to a custom stop list until the real themes appeared.

3. Results & Discussion

The primary results look not that clean, which makes me wonder if it's because I skipped using isalpha() this time. As a result, words attached to commas, periods, or quotation marks may have slipped into the analysis. I noticed unwanted tokens like "mr," even though I tried filtering by word length. I should have tried different code to make it more effectively.

Result LLR

This resulted in positive values for Slave Fiction terms and negative values for Romance Fiction terms. We can cleary see that many words are specific names.

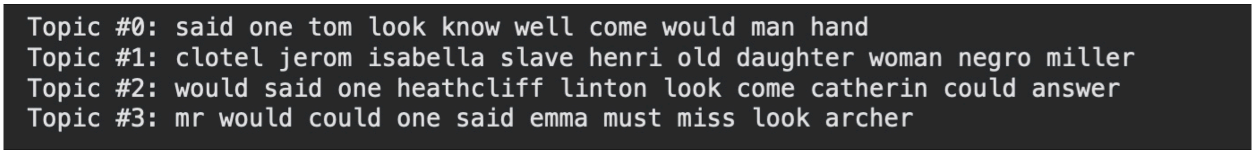


```
   word      llr  freq_slave  freq_romance
  clare 5.953434         374             0
    eva 5.772711         312             0
ophelia 5.753356         306             0
 steven 5.726952         298             0
     ma 5.678997         284             0
    tom 5.594853         785             2
       word       llr  freq_slave  freq_romance
       emma -6.736221           0           864
     archer -6.545774           0           714
heathcliff -6.141008           0           476
     weston -6.080515           0           448
     linton -6.013747           0           419
   catherin -6.011363           0           418
```

I filtered names out after this step, it looked a bit better, but still see different names came up, (at this steps I had an idea that names must be out from my analysis) so I fillter more names out for the next steps.

Result from LDA baseline

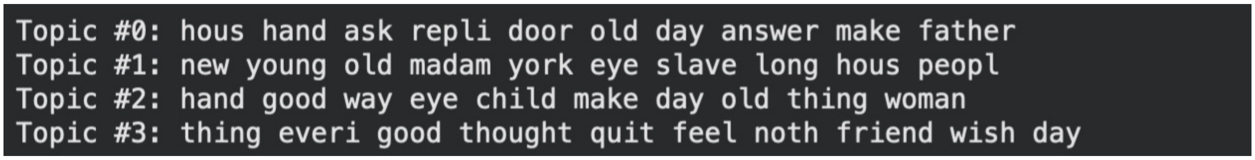
Run 1: The results contain many character names and general words (not yest filter names for baseline), but overall I think they look reasonable. I can clearly see that the topics reflect different books—for example, words like "Emma," "negro," "Tom," and "Linton" appear, which come from four different novels (Topics 0 and 1 from Slave Fiction, and Topics 2 and 3 from Romance Fiction).



```
Topic #0: said one tom look know well come would man hand
Topic #1: clotel jerom isabella slave henri old daughter woman negro miller
Topic #2: would said one heathcliff linton look come catherin could answer
Topic #3: mr would could one said emma must miss look archer
```

Result from LDA Experiments

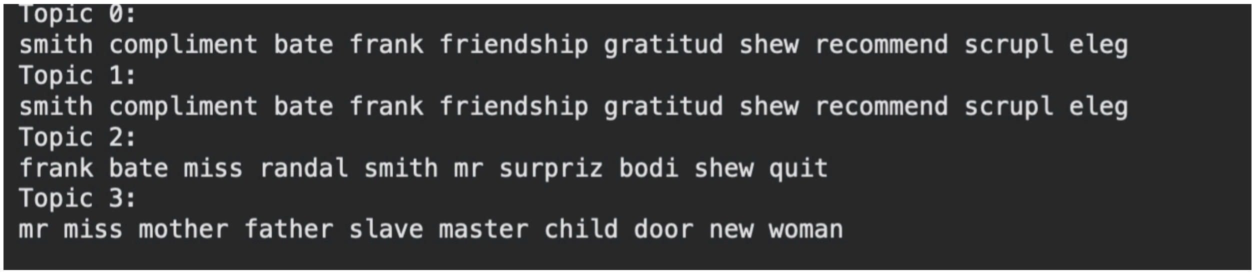
Runs 2–4: Experimental (Run2 =Names Removed, Run3=Fiction "Junk" Removed, and Run4=Combined Filters(2+3)): After removing character names and the fiction-specific "junk" words, clearer thematic patterns began to emerge. In the final result (Run 4), the topics appear to better reflect genre-level distinctions, such as themes related to high society or family life.



```
Topic #0: hous hand ask repli door old day answer make father
Topic #1: new young old madam york eye slave long hous peopl
Topic #2: hand good way eye child make day old thing woman
Topic #3: thing everi good thought quit feel noth friend wish day
```

Results TF-IDF and experiment

In this run, I try with the words that are too rare (specific names), and I don't want words that are too common (generic verbs). I try to see the words in the middle



```
Topic 0:
smith compliment bate frank friendship gratitud shew recommend scrupl eleg
Topic 1:
smith compliment bate frank friendship gratitud shew recommend scrupl eleg
Topic 2:
frank bate miss randal smith mr surpriz bodi shew quit
Topic 3:
mr miss mother father slave master child door new woman
```

During the experimental phase, the model initially separated the genres but still exhibited Partial Collapse. For example, Topics 0 and 1 appeared identical, focusing on general social vocabulary. However, Topic 3 successfully isolated the "Slavery" theme, while Topic 2 captured the "Romance" character names. This result demonstrates that while the thematic signals are present, the model required strict filters (min_df=30 and max_df=0.70) to separate them effectively. (slave have high scores in topic 3)

4. Analysis

1) I frequently observed that the model became trapped by a single dominant pattern, often a specific book's vocabulary(my guess since it looked the same). It is failing to differentiate between the two genres. This was likely due to the noise from character names and generic 1850s narrative descriptions, which acted as a confusing signal for the algorithm.

2) In comparing the techniques, I found that the Log-Likelihood Ratio (LLR) was the most effective tool because it calculates the mathematical "signature" of two known groups. While LDA is typically better for discovering unknown topics in massive datasets, my LDA baseline struggled with "Character Dominance" in this small 6-novel corpus.

   Comparison of Simple Count vs. TF-IDF When comparing Simple Count LDA and TF-IDF, I found that Simple Count occasionally produced more "grounded" results. While TF-IDF (high filter) successfully separated the documents into distinct genres, it did "muted" the most obvious thematic word like "slave."

   Because the word "slave" appeared in nearly every chapter of the Slave Narratives, its IDF score was low, causing the model to treat it as a background stop-word. Instead, the TF-IDF topics relied on higher-scoring, scene-specific words like "market," "prison," and “punish”, “auction”, “escape" to define the genre. In contrast, the LLR analysis—which compares groups rather than document uniqueness, correctly identified "slave" as the single most significant term distinguishing the two corpora.

3) During the iterative testing phase. To resolve the Topic Collapse, I implemented higher frequency cutoffs. By removing words that only appeared in a small number of chapters, I cleared the local noise. This suggests that aggressive pruning was necessary to make the LDA model effective on this specific dataset.

5. Discussion & Personal Reflection

- I notice and I beleive that for fiction especially, preprocessing matters more than the model itself. Cleaning choices can dramatically change LDA behavior.

- I found that my dataset likely caused character names to dominate the topics. Perhaps the dataset is too small; using a larger corpus could dilute the influence of specific names and allow the model to focus on thematic concepts rather than individual plot points.

- I learned the power of stopwords: standard English stop lists aren't enough for literature or fiction. I also learned that there may be no "magic" model—it's more about fine-tuning and adjusting methods to match your research objectives and what you are looking for in the data. Data cleaning is still a crucial step, but it also requires experience and some principles to guide your decisions. Otherwise, you may not know what actions to take.

- For example, when I created additional word filters, I hesitated to remove some words even though they seemed very general, like "see", "would" or "said." These could be considered "fiction junk"—words that carry little meaning—yet it's not always clear what is truly uninformative. To handle this, I tried three different filtering approaches to see how the results would differ.

- This experience makes me want to work with different corpus and run more models. Once you gain enough understanding, you can confidently select and fine-tune the model for your objectives. I want to get to the stage where I know why one approach is superior to another—I have that confidence in my own work, but I am still building it in NLP.

- I wonder what any reason that we will keep a name for analysis in the novels. But in this assignment I dont need it, it is not answer my propose that I want to see words that divide two theme or society, I see some like negro, slave but expected more.

- I admit that I was confused at times, which led me to believe that It is essential to understand the underlying logic of the entire process, specifically that Math (Topics) does not always equal Meaning (Themes). Without understanding the math, it is easy to become confused when the model's logic doesn't match the way a human mind thinks about themes. (Ex. "prison," and "slave" because they appear in the same chapters. To the computer, these are just numbers and probabilities, not "suffering" or "history"). Recognizing this gap allowed me to adjust my preprocessing to better align the statistical results.

- Besides all the topics above, I also learned a few practical things about Python. For example, I learned more about logic, how inputs and outputs are handled, and how Python "sees" or interprets data. I realized that I should avoid putting print statements in different cells, and that using better variable names is important. I lost track a few times because of my own variable choices—using very basic names like X is good at somecases, but descriptive names would have helped though.

- Since I already have the code, I plan to simply change the URLs to different books to see how the results vary. I am interested in trying non-fiction slave narratives, which I have seen on Project Gutenberg.I wanted to use it this time but not sure how to compare with other books. My question is: what should I compare them with? Can I compare fiction with non-fiction, or should I instead compare non-fiction texts across different thematic categories?

AI Clarification & Usage

I used AI as a collaborative assistant to help bridge the gap between my code implementation and the theoretical concepts we learned in class. The AI did suggest lines of code when I was unable to get certain outputs to print. I like when it explain math model to me in a simple words. Overall, It functioned more like a peer for debugging and deeper analysis. I check on general site too when it related to standard process in NLP. Mostly I want to know what nlp coder do in specific case.

AI helped me the most in the first section—setting up the environment—because I was stuck at that stage like getting Python to read files from the web, and combining multiple files into two clear categories. However, once everything was set up properly and organized without errors, it became much easier to complete the rest of the project.

Key examples of my collaboration with AI include:

- Logic Debugging: When I noticed words like "ma" and "me" and ‘mr” in my final results, I asked ai why and ai suggested me to do len > 2 filter, I consulted the AI to understand how noise works, how the filter works which that seem like majors problem here at my dataset
- Methodological Interpretation: I used the AI to help explain the mathematical at TF-IDF, especially the relationship with LDA, to ensure that I compare it in the right way.
- AI is very aggressive, tell me to set TF-IDF pruning very high to force the model toward thematic convergence. I am not yet believing AI at this stage.
- When I consults about the pruning or filters, AI said no standard number, kind of depend. Was that true?
- Check Mathematical Grounding: math logic, heavy consult on this, cos I use the library and I want to do know how they think.
- Iterative Strategy: I asked about general standards in NLP, such as whether it is acceptable to manually filter words. I also wondered whether LDA results might be similar to, or even better than, TF-IDF results. However, I realized that I cannot always be completely sure the answers are correct. It is better to consult reliable sources so that I can verify the information and better understand the theoretical foundations behind these methods.
- It always said “contents may violate” and I belive they don’t like negative word like negro.