Talking More, Saying Less

Corwin Lee

Abstract: This project investigates how narrative structure varies across mediums by comparing the average word counts of movie scripts and books across five genres. Using Python, the program scraped 25 film scripts from Springfield! Springfield! and analyzed 25 public domain books stored as local .txt files. Each text was categorized by genre such as Romance, Sci-Fi, or Horror and processed to compute word totals. The results show a substantial difference in verbosity between the two mediums: books averaged between 77,000 and 140,000 words, while movie scripts ranged from 7,400 to 14,000. Within each medium, genre-specific patterns also emerged, Romance novels contained the most words, while Comedy led film genres in word count. Conversely, Action and Horror films were the most condensed. These findings demonstrate how both genre and medium influence storytelling strategies, offering insights into how narrative density adapts to the constraints and strengths of different formats.

Specs: I used Python 3.11, the requests and re libraries for web scraping and text parsing, and Matplotlib for data visualization. The project was coded locally using VS Code and executed on a Replit environment. All movie script data was scraped from Springfield! Springfield!, and book texts were processed from local .txt files downloaded from public domain sources such as Project Gutenberg.

Website: https://github.com/CKWlee/BookMovieWordComparison

Contact: For more information, please contact Corwin Lee at ckwlee@umass.edu

Corwin Lee

Professor Stephen Harris

English491DS: Data Science In Humanities

May 1, 2025

Talking More, Saying Less

Introduction

Stories can look completely different depending on how they're told. A book might take hundreds of pages to build a world or explore a character's thoughts, while a movie has to get the point across in just a couple of hours, often through visuals and dialogue. This project looks at one basic but telling difference: how many words different genres use in books versus movies.

Using a Python script, I gathered 25 movie scripts and 25 books, organized them by genre, and calculated their average word counts. By comparing genres like Comedy, Sci-Fi, and Horror across these two formats, I wanted to see which ones tend to be more wordy, and how storytelling styles shift depending on the medium. The goal is to get a better sense of how genres adapt when the format changes and what that might say about how we consume stories.

Data Sources

Movie scripts were collected from the online repository Springfield! Springfield!, a public archive of screenplays and transcripts. A total of 25 scripts were scraped, covering 5 genres:

- Comedy: *Superbad*, *Step Brothers*, *The Hangover*, *Anchorman*, *Bridesmaids*

- Sci-Fi: *Inception*, *Interstellar*, *The Matrix*, *Blade Runner*, *Arrival*
- Drama: *Forrest Gump*, *Shawshank Redemption*, *The Godfather*, *A Beautiful Mind*, *The Pursuit of Happiness*
- Action: *Die Hard*, *Mad Max: Fury Road*, *John Wick*, *Gladiator*, *The Dark Knight*
- Horror: *Get Out*, *The Shining*, *Hereditary*, *The Conjuring*, *Scream*

Book texts were obtained as .txt files, primarily from public domain sources such as Project Gutenberg. 25 books were processed, grouped into:

- Horror: *Dracula*, *Frankenstein*, *The Phantom of the Opera*, *The Beetle*, *The Trial*
- Romance: *Pride and Prejudice*, *Jane Eyre*, *Emma*, *Sense and Sensibility*, *Persuasion*
- Science Fiction: *War of the Worlds*, *The Invisible Man*, *We*, *Journey to the Center of the Earth*, *Twenty Thousand Leagues Under the Sea*
- Mystery: *The Hound of the Baskervilles*, *The Mystery of the Yellow Room*, *The Mysterious Affair at Styles*, *The Woman in White*, *Daredevil*
- Adventure: *Moby Dick*, *Treasure Island*, *Tom Sawyer*, *Huckleberry Finn*, *Robinson Crusoe*

Methodology

1. Data Collection and Preprocessing

- Movie scripts were fetched using requests and parsed using re (regular expressions) to extract text content from the HTML “<div>” containing the screenplay.
- Book files were accessed locally and read into memory with UTF-8 encoding.

2. Word Count Calculation

- Each text was tokenized by matching all contiguous word characters using the pattern “\b\w+\b”.
- Word totals were calculated for each script or book.
- Only successful extractions (non-zero word count) were included in average calculations.

3. Aggregation and Genre Analysis

- Word counts were grouped by genre.
- Averages were computed across each genre group for both books and movies.
- Results were visualized using matplotlib, producing two separate bar charts:
  - ○ Average Word Count by Movie Genre
  - ○ Average Word Count by Book Genre

Limits and Constraints

- Completeness: It was assumed that all scripts and books were full texts and representative of the genre.
- Word Count as Metric: Word count does not account for content quality, pacing, or subtext but serves as a useful quantitative proxy.
- Genre Categorization: Some overlap exists between genres, but each text was assigned to only one genre for clarity.
- Sample Size: Five texts per genre is modest but sufficient to identify broad patterns.

- Constitutions of a novel: The difference between novella and a novel was less defined in the times of classics.

Results

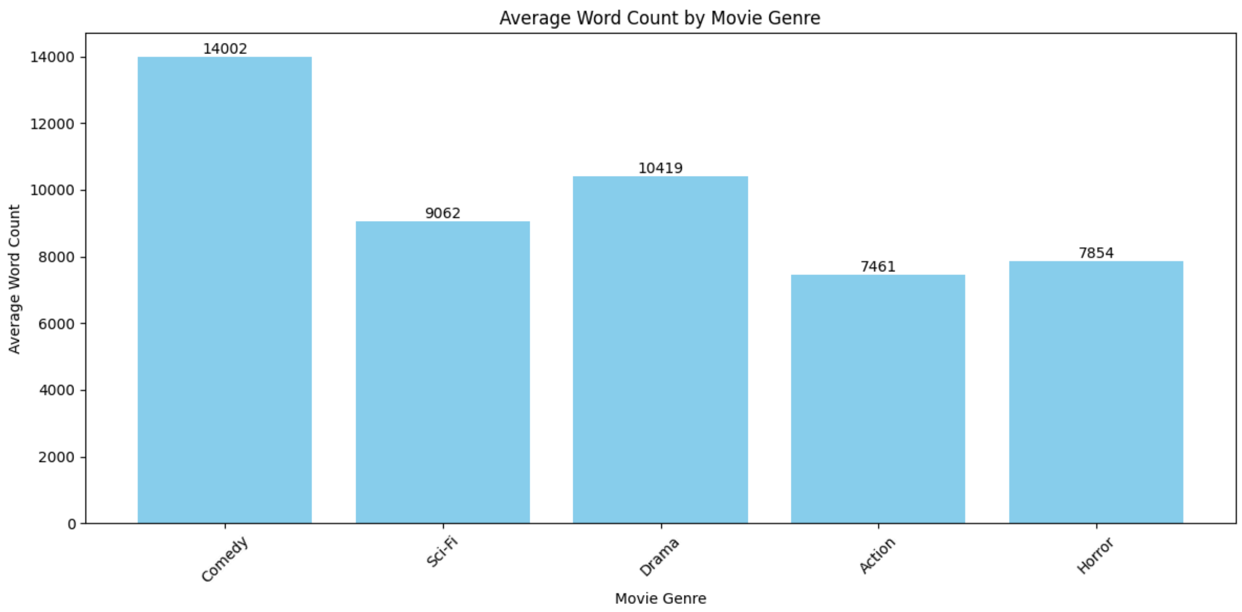
Movie Genre Word Count (Average)

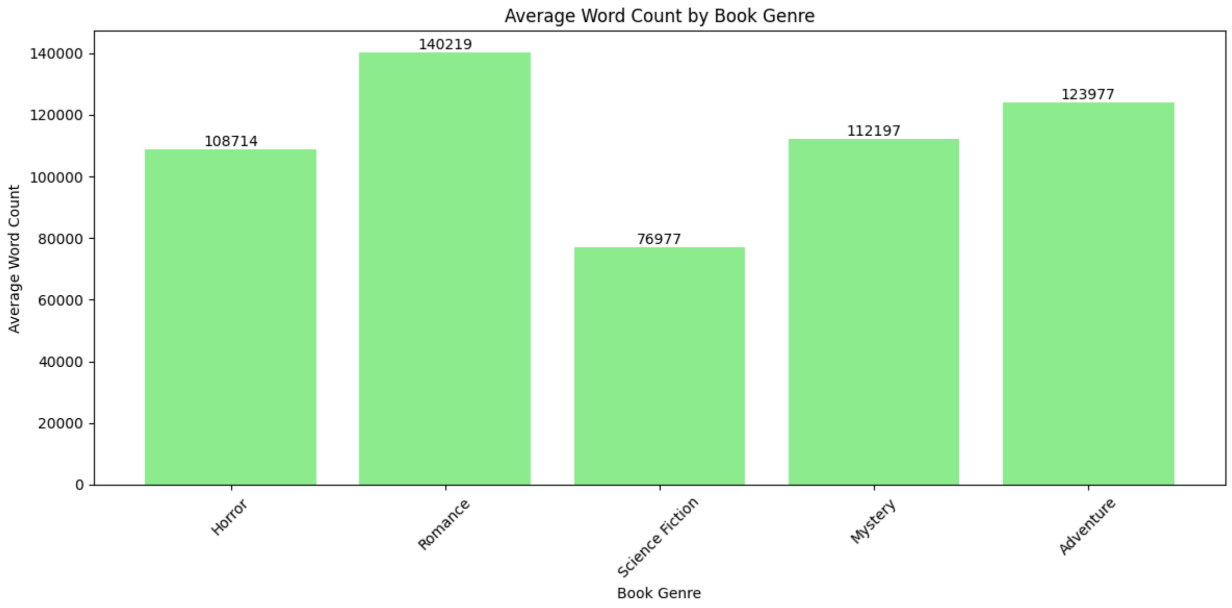
Genre	Average Words
Comedy	14,002
Science Fiction	9,062
Drama	10,419
Action	7,461
Horror	7,854

Book Genre Word Count (Average)

Genre	Average Words
Horror	108,714
Romance	140,219
Science Fiction	76,977
Mystery	112,197
Adventure	123,977

Data Visualization





Analysis

The results reveal consistent and significant differences in average word counts between books and movies, as well as noteworthy genre-specific trends within each medium.

Books naturally contain more words than movies, with averages ranging from 76,977 words (Science Fiction) to 140,219 words (Romance). This disparity reflects the structural freedom that novels have to delve into detailed settings, character arcs, and elaborate plots. Among literary genres, Romance stands out as the most verbose, likely due to its emphasis on emotional complexity, internal monologue, and extended character arcs. Adventure (123,977 words) and Mystery (112,197 words) also show high word counts, supporting the idea that suspense and world-building often require substantial narrative worldbuilding. Conversely, Science Fiction averages the fewest words in books (76,977), which may be due to the selection of shorter, more conceptual novels in the sample set rather than a universal trend.

In the film medium, the average word counts are considerably lower, ranging from 7,461 (Action) to 14,002 (Comedy). This is expected, as film scripts are limited by runtime and rely heavily on visual storytelling. comedy has the highest average word count, suggesting that this genre leans heavily on fast-paced dialogue, rapid exchanges, and verbal punchlines to drive humor. Drama (10,419 words) also ranks high, as it often relies on dialogue to reveal character depth and emotional stakes. In contrast, Action and Horror show the lowest word counts, at 7,461 and 7,854 respectively, reinforcing the idea that these genres often use visuals, pacing, and sound design to create tension and engagement.

Science Fiction, despite its complexity, lands in the middle of both charts: third lowest in books and second lowest in movies. This suggests that while the genre involves detailed world-building, it may be because of the usage of dense language or reliance on the audience's familiarity with tropes.

These findings support the hypothesis that medium and genre jointly shape narrative structure. Genres with an emphasis on internal development or verbal wit like Romance in books or Comedy in film tend to be wordier. Meanwhile, genres that depend on spectacle like Action or Horror lean on nonverbal storytelling methods. Word count, while a simplistic metric, serves as a useful proxy for understanding how different genres engage their audiences and how narrative strategies adapt across mediums.

Conclusion

This project highlights how storytelling is deeply influenced by the medium in which it unfolds. Books, with their freedom from time constraints and visual limitations, allow for expansive narratives that often exceed 100,000 words. In contrast, movie scripts are bound by runtime and rely on visual and auditory cues, resulting in leaner, dialogue-driven texts averaging around 10,000 words.

Across both media, genre plays a pivotal role in shaping narrative density. Romance novels emerged as the most verbose, reflecting the genre's focus on emotional introspection and relationship development. Comedy scripts, on the other hand, proved to be the wordiest among films likely due to their reliance on quick, snappy dialogue and verbal humor. Meanwhile, genres like Action and Horror consistently favored brevity in film form, relying more on spectacle and atmosphere than on language.

While word count alone cannot capture the full richness of a story, it offers valuable insight into the structural and stylistic conventions of different genres. This comparison highlights how narrative content is shaped not only by creative choices but also by medium-specific limitations and audience expectations.

Future work could explore additional variables such as sentence complexity, dialogue-to-description ratio, or pacing. It might also be valuable to apply similar analysis to emerging formats like streaming miniseries or interactive media, where the boundaries of storytelling continue to evolve.