CompSci 4NL3
Assignment 2
Harshit Sharma
400349986

Corpus Analysis Report

1. Dataset

The dataset used for this analysis consists of two movie scripts: *The Avengers* and *Drag Me to Hell*. These were chosen to explore linguistic patterns in different genres—superhero action and horror and also because I am a huge Marvel fan and I really love the first Avengers movie and Drag Me To Hell was the first horror movie I saw with my brother and I found it scary but funny because of the goat who is the devil incarnate dancing. This comparison allows for an examination of how language usage differs between genres and how topic modeling can capture thematic variations.

Data Collection

The scripts were sourced as plain text files and processed using a document-splitting approach. Each script was divided into 100 segments to ensure a balanced dataset.

Dataset Overview

Category	Number of Documents	Avg. Tokens per Document
Superheroes	100	~200
Horror	100	~200

This approach maintains uniformity across both categories, ensuring that differences observed in later analyses are genre-based rather than length-based.

2. Methodology

Preprocessing Steps

1. **Text Cleaning:**
    - Lowercased text to maintain consistency.
    - Removed punctuation and special characters.
    - Eliminated URLs and numbers.
2. **Tokenization:**
    - Used manual splitting since external NLP libraries were not allowed.
3. **Stopword Removal:**
    - Removed common English stopwords.
4. **Character Name Filtering:**
    - Excluded major character names (e.g., "Tony," "Loki," "Natasha") to prevent genre bias.
5. **Lemmatization and Stemming:**
    - Lemmatization was used to normalize words.
    - Stemming was an optional experiment but was not applied in the final version due to distortions in readability.

Analysis Performed

1. **Bag-of-Words Representation:**
    - Converted the cleaned text into a frequency matrix using `CountVectorizer`.
2. **Naïve Bayes Log-Likelihood Analysis:**
    - Calculated the most distinctive words for each genre based on log-likelihood ratios.
3. **Topic Modeling with LDA:**
    - Identified key topics within each genre using Gensim's LDA model.
    - Defined five meaningful topic labels per genre.

3. Results and Analysis

Top Words per Category

Using log-likelihood ratios, the most distinctive words for each genre were identified:

Superhero Genre (Top 10 Words)

Word	Log-Likelihood Ratio
agent	4.823533
man	1.807733
look	0.621008
nick	4.642535
iron	2.818515
day	1.234748
int	0.479598
coulson	4.328042
ext	1.059016
captain	4.317113

Horror Genre (Top 10 Words)

Word	Log-Likelihood Ratio

stephanie	6.535868
ray	4.130462
mr	4.798917
stephanies	4.671083
old	2.617284
woman	2.918326
back	-0.042642
look	-0.621008
jack	4.236798
hand	0.379849

Topic Modeling Results

Using LDA with five topics per genre:

Superhero Topics

Topic Label	Top Words (with probabilities)

Superpowers & Science	captain (0.0254), chitauri (0.0246), america (0.0225)
Combat & Military	iron (0.0409), ext (0.0233), captain (0.0174)
Leadership & Strategy	chitauri (0.0431), iron (0.0408), hand (0.0229)
Technology & Weapons	nick (0.0378), world (0.0326), security (0.0291)
Tactical Planning	agent (0.0393), coulson (0.0282), nick (0.0214)

Horror Topics

Topic Label	Top Words (with probabilities)
Curses & Rituals	san (0.0411), dena (0.0362), milo (0.0226)
Possession & Spirits	ilenka (0.0343), door (0.0333), house (0.0182)
Paranormal Encounters	ray (0.1813), dont (0.0160), train (0.0156)
Psychological Terror	old (0.0517), mr (0.0501), woman (0.0410)
Demonic Entities	jack (0.0274), mr (0.0248), bank (0.0169)

Top Topics per Category

Superheroes

1. Tactical Planning (42.7%)
2. Combat & Military (19.9%)
3. Superpowers & Science (16.4%)
4. Technology & Weapons (12.8%)
5. Leadership & Strategy (8.1%)

Horror

1. Demonic Entities (36.9%)
2. Psychological Terror (18.2%)
3. Curses & Rituals (17.9%)
4. Paranormal Encounters (17.5%)
5. Possession & Spirits (9.3%)

4. Discussion

4.1 Findings About the Dataset

From this analysis, clear linguistic patterns emerged based on genre. Superhero scripts emphasize action, strategy, and technological advancements, while horror scripts focus on psychological dread, supernatural elements, and suspenseful imagery. These findings reinforce how storytelling differs structurally between these genres.

4.2 Lessons Learned

- **Data Preprocessing is Crucial:** The filtering of character names was essential to prevent misleading results.
- **Bag-of-Words vs. Topic Modeling:** While BoW captures frequency, LDA provides deeper insights into thematic structures.
- **Experimentation with Text Normalization:** Using both stemming and lemmatization revealed that lemmatization preserved semantic integrity better.
- **Limitations:**
    - The dataset was limited to two scripts, which may not fully generalize across entire genres.
    - More sophisticated models, such as BERT-based topic modeling, could provide further insights.

Conclusion

This project successfully demonstrated how computational text analysis can extract meaningful insights from different genres. Future work could involve expanding the dataset and experimenting with deep learning techniques for improved classification and topic modeling.