Virendra Jethra (jethrav)

4NL3 Homework 2 Report

1. Data:

   For this project, I used five books from the Harry Potter series and five books from the Percy Jackson series, which are two distinct categories. The books were split into individual chapters, with Harry Potter containing 132 chapters/documents and Percy Jackson containing 105 chapters/documents, for a total of 237 documents.

   I chose this dataset because both series are highly popular fantasy novels with rich world-building, unique themes, and distinct storytelling styles. So analyzing them would provide me insight into the differences in language, themes, and narrative structures between the two series. I collected this data from these online resources (Harry Potter, Percy Jackson) converted it into a plain text file and used a python script to split it into chapters. I preprocessed the corpus using my last assignment preprocessing function with little modification.

Category	**Number of Documents**	**Average Tokens per Document**	**Total Tokens**
Harry Potter	132	2,790	368,284
Percy Jackson	105	1,973	207,162
Total	**237**	**4,763**	**575,446**

2. Methodology:
   a. To begin, I installed all the necessary libraries for this assignment, including nltk, gensim, pyLDAvis, and pyLDAvis.gensim_models, numpy, and pandas. Next, I preprocessed the corpus using a modified version of my previous assignment's function. Instead of using the nltk library for stopwords, I used a GitHub post that provided a more relevant stopword list for the text, helping to remove unnecessary information. Additionally, I removed character names to extract more meaningful insights. For preprocessing, I applied lowercasing and stopword removal, along with the removal of specific character names, while excluding stemming and lemmatization as they were not essential for this analysis. Since keeping words in their natural form helped the LDA model learn better topic distributions without unnecessary alterations.

After preprocessing, I created a function to load the corpus using the OS file system. This function processes each document and merges them into a single list. It also acts as a bag-of-words (BoW) generator by extending the list as a dictionary and returning tokens with their respective counts.

To compute the LLR, I utilized the list of dictionaries generated from the load_corpus function. I calculated the total number of tokens and created a vocabulary list using dictionary keys and values. Then, I computed the log-likelihood ratio and returned a sorted list of significant words.

For LDA, I used the gensim library to convert tokens into word IDs and transformed each document into a BoW representation. I then trained the LDA model using gensim and visualized the results with pyLDAvis.

To analyze the topics, I implemented a function that extracts the top 25 words for each topic and presents them in a structured table. Finally, I created a function to compute the average topic distribution across all documents within each category, helping to identify the most prominent themes in Harry Potter and Percy Jackson.

3.

a. Here are my results for Naive Bayes (LLR), the top 10 words sorted by their log-likelihood ratios for each class with their score:

Harry Potter's Top Words	**Percy Jackson's Top Words**
wand: 6.26	kronos: 6.43
malfoy: 6.15	zo: 6.37
hogwarts: 5.91	clarisse: 6.29
mcgonagall: 5.78	rachel: 6.25
neville: 5.76	mom: 6.06
gryffindor: 5.64	olympus: 5.89
ginny: 5.45	bianca: 5.83
ministry: 5.42	artemis: 5.76
vernon: 5.40	jackson: 5.69
fudge: 5.39	guy: 5.66

The main takeaway from computing the LLR was that identifies keywords distinguishing the Harry Potter category (wizardry, Hogwarts, magical figures) from the Percy Jackson category (Greek mythology, gods, demi-gods). High-scoring words like "wand" and "kronos" confirm the differences, proving LLR's effectiveness in classifying texts based on vocabulary.

b. Here are my results for Topic Modelling (LDA), the top 25 terms for that topic, sorted by their probability of belonging to that topic:

Words	Topic 1: Dark Magic & Villains	Topic 2: Heroes & Hunters	Topic 3: School Life	Topic 4: Gods & Divine Beings	Topic 5: Water & Nature	Topic 6: Professors & Authority	Topic 7: Family & Demigod Origins	Topic 8: Cyclops & Sea Creatures	Topic 9: Ministry of Magic & Bureaucracy	Topic 10: Competitions & Challenges
1	frank (0.0141)	didnt (0.0072)	malfoy (0.0072)	didnt (0.0072)	ter (0.0052)	professor (0.0074)	zo (0.0078)	clarisse (0.0073)	looked (0.0064)	looked (0.0056)
2	wormtail (0.0060)	looked (0.0063)	looked (0.0054)	looked (0.0069)	asked (0.0046)	looked (0.0062)	looked (0.0069)	polyphemus (0.0050)	didnt (0.0053)	time (0.0044)
3	voice (0.0036)	zo (0.0054)	dont (0.0054)	dont (0.0059)	looked (0.0038)	dont (0.0046)	didnt (0.0065)	didnt (0.0045)	black (0.0048)	didnt (0.0043)
4	riddles (0.0029)	time (0.0043)	didnt (0.0050)	time (0.0058)	dont (0.0037)	face (0.0043)	dont (0.0063)	looked (0.0045)	eyes (0.0047)	bagman (0.0042)
5	lord (0.0027)	dont (0.0043)	time (0.0046)	eyes (0.0050)	told (0.0037)	eyes (0.0042)	bianca (0.0053)	dont (0.0041)	dont (0.0047)	water (0.0035)
6	man (0.0024)	eyes (0.0036)	eyes (0.0039)	kronos (0.0042)	face (0.0033)	time (0.0042)	eyes (0.0053)	sheep (0.0037)	time (0.0040)	dont (0.0033)
7	dont (0.0023)	asked (0.0035)	lockhart (0.0037)	gods (0.0041)	didnt (0.0032)	voice (0.0040)	told (0.0047)	time (0.0036)	face (0.0036)	eyes (0.0031)
8	police (0.0023)	face (0.0033)	told (0.0030)	told (0.0038)	didnt (0.0032)	didnt (0.0039)	time (0.0047)	cyclops (0.0034)	told (0.0034)	head (0.0031)
9	house (0.0021)	half (0.0033)	dark (0.0030)	asked (0.0037)	eyes (0.0032)	head (0.0038)	mom (0.0035)	told (0.0033)	head (0.0032)	told (0.0030)
10	cold (0.0019)	turned (0.0031)	face (0.0030)	rachel (0.0035)	time (0.0031)	door (0.0038)	good (0.0031)	voice (0.0032)	voice (0.0031)	felt (0.0029)

11	riddle (0.0019)	told (0.0031)	people (0.0030)	face (0.0034)	water (0.0031)	room (0.0036)	thought (0.0031)	ship (0.0030)	turned (0.0029)	wand (0.0028)
12	stick (0.0018)	camp (0.0029)	heard (0.0030)	half (0.0034)	good (0.0030)	wand (0.0034)	apollo (0.0030)	water (0.0030)	stan (0.0028)	good (0.0027)
13	thought (0.0017)	big (0.0028)	fred (0.0029)	camp (0.0034)	head (0.0029)	thought (0.0032)	asked (0.0030)	face (0.0029)	fudge (0.0024)	cedric (0.0027)
14	didnt (0.0017)	head (0.0028)	neville (0.0028)	good (0.0032)	turned (0.0024)	hand (0.0029)	half (0.0029)	sea (0.0029)	asked (0.0023)	krum (0.0026)
15	looked (0.0017)	thought (0.0027)	good (0.0028)	thought (0.0031)	long (0.0023)	long (0.0028)	gabe (0.0029)	head (0.0028)	long (0.0023)	turned (0.0025)
16	village (0.0016)	knew (0.0026)	wand (0.0027)	turned (0.0030)	thought (0.0023)	turned (0.0028)	knew (0.0029)	good (0.0027)	artemis (0.0023)	face (0.0024)
17	room (0.0015)	artemis (0.0025)	george (0.0026)	knew (0.0030)	river (0.0023)	good (0.0028)	camp (0.0028)	knew (0.0027)	good (0.0022)	long (0.0024)
18	told (0.0015)	good (0.0024)	head (0.0026)	head (0.0029)	yeh (0.0023)	told (0.0028)	monster (0.0027)	eyes (0.0026)	pettigrew (0.0022)	quirrell (0.0023)
19	time (0.0014)	long (0.0023)	voice (0.0026)	long (0.0028)	nick (0.0022)	black (0.0028)	long (0.0026)	geryon (0.0026)	hand (0.0022)	thought (0.0023)
20	eyes (0.0014)	voice (0.0023)	black (0.0025)	voice (0.0028)	ive (0.0022)	fred (0.0028)	school (0.0025)	half (0.0025)	started (0.0021)	knew (0.0023)
21	door (0.0013)	yelled (0.0022)	turned (0.0025)	hand (0.0027)	ill (0.0021)	hogwarts (0.0025)	turned (0.0024)	monster (0.0025)	hands (0.0020)	voice (0.0022)
22	hangleton (0.0012)	sword (0.0022)	wood (0.0024)	sword (0.0026)	voice (0.0021)	dark (0.0025)	wanted (0.0021)	father (0.0024)	yelled (0.0020)	tent (0.0022)
23	bryce (0.0012)	monster (0.0021)	thought (0.0024)	big (0.0026)	yeah (0.0021)	people (0.0025)	artemis (0.0021)	sword (0.0023)	knew (0.0020)	hand (0.0021)
24	nagini (0.0012)	hunters (0.0021)	gryffindor (0.0023)	monsters (0.0025)	headless (0.0020)	malfoy (0.0025)	face (0.0021)	yelled (0.0023)	bus (0.0020)	moody (0.0020)
25	head (0.0011)	thorn (0.0020)	door (0.0023)	cabin (0.0023)	black (0.0020)	heard (0.0024)	metal (0.0020)	monsters (0.0023)	wanted (0.0020)	great (0.0020)

The top 5 topics for each category are:

Harry Potter Top Topics	Percy Jackson's Top Topics
(Professors & Authority, 0.8019474)	(Gods & Divine Beings, 0.50165504)
(School Life, 0.118501134)	(Heroes & Hunters, 0.18685402)
(Competitions & Challenges, 0.031639554)	(Family & Demigod Origins, 0.101988845)
'Water & Nature, 0.01775591)	(Cyclops & Sea Creatures, 0.0862686)
(Ministry of Magic & Bureaucracy, 0.016593348)	(Ministry of Magic & Bureaucracy, 0.06081561)

The main takeaway from this LDA topic modelling evaluation is that it effectively distinguishes topics in both categories. Harry Potter focuses on school life, professors, and magical bureaucracy, while Percy Jackson highlights gods, heroes, and mythological creatures. Given the content of both series, it makes sense that these topics were identified by LDA. The model successfully groups related words, revealing key themes in each series. The slight overlap, such as bureaucracy, suggests shared structural elements. Overall, LDA is useful for automated text categorization, topic discovery, and analyzing large text collections by identifying meaningful word associations.

c. For the experimental part of the assignment, I did a minimum length of 3 for the text normalization which changes the LDA analysis. Here are the top 5 topics from each category:

Harry Potter Top Topics	Percy Jackson's Top Topics
('School Life', 0.69597137)	(Heroes & Hunters, 0.70304203)
(Competitions & Challenges, 0.10778512)	(Gods & Divine Beings, 0.17362456)
(Professors & Authority, 0.075683475)	(Water & Nature, 0.066963024)
(Cyclops & Sea Creatures, 0.046250407)	(Professors & Authority, 0.030988272)

(Family & Demigod Origins, 0.03152087)	(Family & Demigod Origins, 0.0128959175)

As we can see the results are different from the original, as filtering out words shorter than three characters improved LDA topic modelling by reducing noise and enhancing topic clarity. In the Harry Potter category, "School Life" emerges as the dominant theme instead of "Professors & Authority," providing a more balanced representation. For Percy Jackson, "Heroes & Hunters" becomes the primary focus, highlighting action and adventure. Removing short, common words like “is” and “to” which reduces topic inflation and improves coherence.

For the other variation, I did the bag-of-words in TF-IDF. Here is the result for Naive Bayes (LLR), the top 10 words sorted by their log-likelihood ratios using the TF-IDF:

Harry Potter's Top Words	**Percy Jackson's Top Words**
professor: 2.26	zo: 1.86
malfoy: 1.89	rachel: 1.69
wand: 1.71	clarisse: 1.63
fred: 1.7	kronos: 1.57
mcgonagall: 1.66	gods: 1.48
neville: 1.65	bianca: 1.37
vernon: 1.63	camp: 1.32
moody: 1.57	mom: 1.31
hogwarts: 1.52	artemis: 1.3
fudge: 1.49	ares: 1.24

As we can see using TF-IDF and count-based BoW in LLR differ in how they weigh term importance. Count-based BoW highlights high-frequency words, emphasizing the raw popularity like "wand" and "hogwarts" for Harry Potter, "kronos" and "olympus" for Percy Jackson. However, TF-IDF reduces the weight of common terms and prioritizes category-specific,

unique words like "professor" and "fudge" for Harry Potter, "zo" and "artemis" for Percy Jackson. This makes TF-IDF better at capturing distinct themes and reducing noise from overrepresented terms. Both methods provide meaningful insights, but TF-IDF enhances contrasts between categories by focusing on uniqueness.

4.

a. By analyzing the Harry Potter and Percy Jackson series, I found that each has distinct themes and focuses. Even though both series follow a hero's quest, they explore vastly different themes and worlds. Harry Potter emphasizes school life, with words like "professor," "hogwarts," and "wand" highlighting its academic and magical setting. It also touches on authority figures and competitions. In contrast, Percy Jackson focuses more on heroes, gods, and mythical adventures, with terms like "kronos," "artemis," and "camp" reflecting its mythological world. Additionally, since I removed character names, these results highlight the overarching themes rather than individual characters' prominence. Overall these results show how any corpus can be simplified into its key terms and themes to describe its topics.

b. Through this assignment, I gained a deeper understanding of LLR and LDA models, particularly the LDA model, which was initially challenging as I had no prior experience with the Gensim and pyLDAvis libraries. Even though I was able to get the code through the tutorial, I invested time in researching each argument through documentation and tools like ChatGPT to understand how it can affect the performance of the corpus. Initially, my topic modelling results were unclear due to incorrect arguments in the LDA model, but after fine-tuning them, the topics became much more meaningful and aligned with the themes of my datasets. Additionally, working with TF-IDF for the BoW representation was difficult at first, as I was unfamiliar with it. However, with guidance from tutorials and assistance from ChatGPT, I was able to understand its implementation and successfully integrate it into my code.

Generative AI:

I used the ChatGPT model to ask about the arguments in the LDA models and to implement TF-IDF for the BoW. The carbon footprint is 0.46kg of CO_2 (model: ChatGPT, Hardware: GTX 1660 Ti, Time Used: 0.5h, Provider: Google Cloud Platform, Region of Compute: Us-east1).