Hosanna Root

DATA 340, Fall 2023

Final Project Report

12/25/2023

Abstract

In the realm of contemporary storytelling, the integration of Artificial Intelligence (AI) poses a fascinating frontier. This report presents an innovative exploratory analysis conducted by Hosanna Root for the DATA 340 course in Fall 2023, focusing on the intricacies of classic literature to enhance AI's narrative capabilities. Inspired by the AI Dungeon app, this project aims to craft AI-generated stories that parallel the depth and allure of traditional literary works.

The research began by selecting a corpus of classic literature from Project Gutenberg, comprising five fiction genres: fantasy, romance, mystery, horror, and science fiction. Ten high-rated stories from each genre were chosen based on their public domain status and popularity on platforms like Goodreads, Google, and Amazon Books. These stories underwent meticulous preprocessing to tailor them for detailed Natural Language Processing (NLP) analysis.

Using nltk and spacy libraries, the analysis involved identifying frequently used words, sentence structures, and parts of speech, while TextBlob and Spacy were employed to visualize sentiment polarity and subjectivity. Lexical diversity, sentence length, passive sentence frequency, and Flesch reading scores were also assessed. Some key challenges were distinguishing character interactions and dialogues from non-human textual elements and solving memory and model processing issues.

The project took a novel turn by applying analytical insights to guide genre-specific story generation using Llama-2 through Replicate. Although budgetary constraints limited the scope, initial attempts using GPT-2 and the analysis of shorter Llama-2 generated stories provided valuable insights. The project underscores the potential of combining NLP analysis of classic literature with advanced AI models to revolutionize narrative generation, albeit acknowledging the challenges in handling large text files and the need for more robust computational resources.

This report not only contributes to the field of AI-assisted storytelling but also opens avenues for further research in utilizing NLP for genre-specific narrative enhancement. The findings, accompanied by visual representations, underscore the promise and challenges in this emerging interdisciplinary field.

Introduction

This project represents an exploratory analysis aimed at deciphering the quintessential elements that constitute a compelling story. Additionally, it seeks to utilize these insights to guide the text-generation capabilities of an AI, with the end goal of crafting narratives that rival the quality of the classic literature dataset I examined. My inspiration stems from AI Dungeon, an app launched in 2019, which harnesses Dungeons and Dragons-like story presets and leverages OpenAI's text-generation through user inputs. The app, highly engaging in its free version, prompted me to delve into the intricacies of story generation, aspiring to reach the narrative heights of time-honored fictional works.

Process

To delineate what characterizes 'classic' and 'exemplary' literature, I accessed Project Gutenberg, primarily to circumvent copyright constraints. I selected five fiction genres: fantasy, romance, mystery, horror, and science fiction. From each category, I chose ten stories, ensuring they held a rating of 3.8 or higher on various platforms such as Goodreads, Google, and Amazon Books, and confirmed their public domain status. After downloading the text files, I edited out any Project Gutenberg-specific content to refine the data for analysis and organized them into genre-specific folders.

The initial phase of my Natural Language Processing (NLP) process involved using Jupyter Notebook in an virtual environment python kernel, utilizing nltk and spacy libraries to identify frequently used words, meticulously filtering out articles, quotations, apostrophes, and punctuation. I then examined sentence structures and parts of speech. Utilizing TextBlob, I visualized sentiment polarity (identifying positive and negative connotations) and sentiment subjectivity (distinguishing factual from subjective content). Spacy facilitated the extraction of prevalent entities (places, people, objects) and their occurrence rates within each text. Further analysis using nltk included assessing lexical diversity(usage of unique words or phrases in a text), average sentence length, passive sentence frequency (passive voice usage), and Flesch reading scores (how easy is it to read said story), to gauge narrative accessibility.

A challenging aspect was discerning character interactions and dialogues, particularly with non-human entities like chapter titles and Project Gutenberg headers being erroneously identified as characters, and places or even beginnings of dialogue. I was able to get a clear,

comprehensive list of character interactions by ratio and name, which I visualized using a node graph. The final analytical step involved employing gensim and nltk for Latent Dirichlet Allocation (LDA) to group the most frequently used words across texts into topics, which, while insightful, was unhelpful as the grouped words were too vague to identify what topics they were sorted into.

Notably, while there exists substantial NLP research in story genre analysis, this is not the focus of my project, as using the result of these findings is key to understanding how to prompt an AI model to generate a story of the caliber I have analyzed. Therefore, I ventured to utilize Llama-2-7b-chat model (access obtained through Meta) through Replicate. Although constrained by budgetary limits, this approach enabled me to discern effective parts of speech for distinct genres. I used the findings from my NLP of the stories to help me prompt a very short story (800 characters) for each genre (see image below). I attempted to use GPT-2, but the results were not worth pursuing.

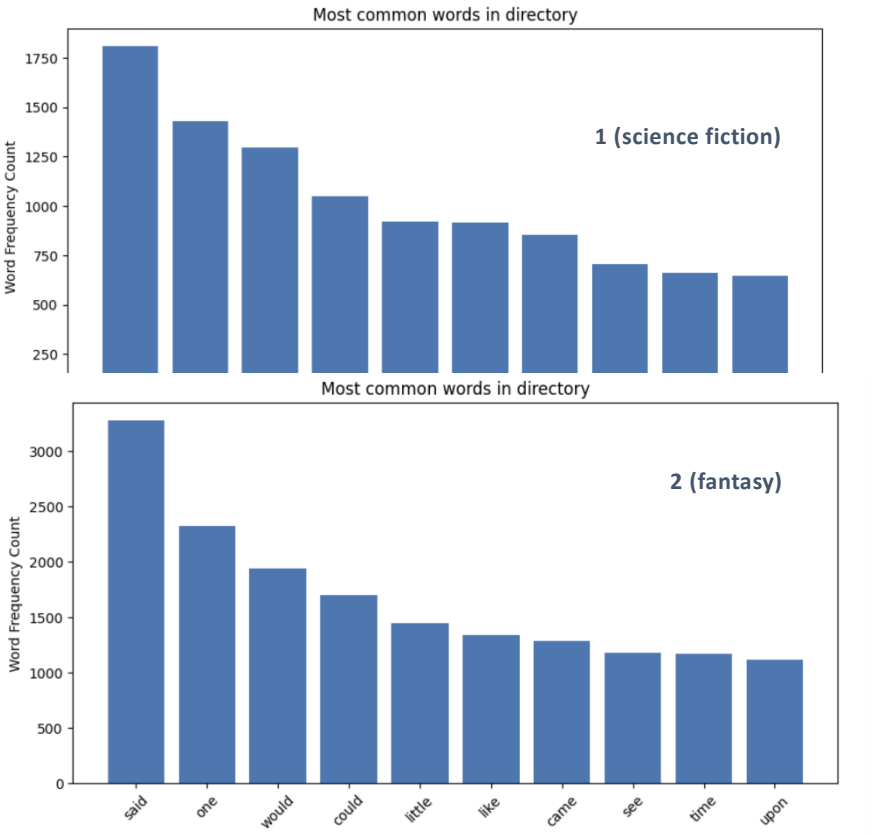
```
pip install replicate

export REPLICATE_API_TOKEN=<paste-your-token-here>

import replicate

output = replicate.run(
  "meta/llama-2-7b-chat
      :13c3cdee13ee059ab779f0291d29054dab00a47dad8261375654de5540165fb0",
  input={
    "debug": False,
    "top_k": -1,
    "top_p": 0.95,
    "prompt": "Generate a fantasy story with short sentences, passive voice, and
        easy to understand dialogue between characters. Use older vocabulary
        style words. Use the words 'one', 'would', 'could', 'little', 'like,
        'said', 'came', 'see', 'time', and 'upon' frequently.",
    "temperature": 0.7,
    "system_prompt": "Intake NLP instructions, and generate a text story based on
        the instructions provided. Do not repeat the instructions in your output
        .",
    "max_new_tokens": 800,
    "min_new_tokens": -1,
    "repetition_penalty": 1.15
  }
)
print(output)
```

The Llama-2 responses were quite short compared to the txt files as they were 800 characters long, so the NLP analysis was lacking compared to the classic stories, especially for character interactions and dialogue. I ran into issues with my device crashing multiple times due to the sheer load of working with pipeline and transformers (from the HuggingFace model hub) on Google Colab's free T4 GPU, as well as generating basic NLP tasks on texts that were very large. If I can work with paid models of Llama-2 or other chat models, and higher GPU capacity in the future, I will be better able to perform model analysis on story generation and prompt engineering with longer length outputs.

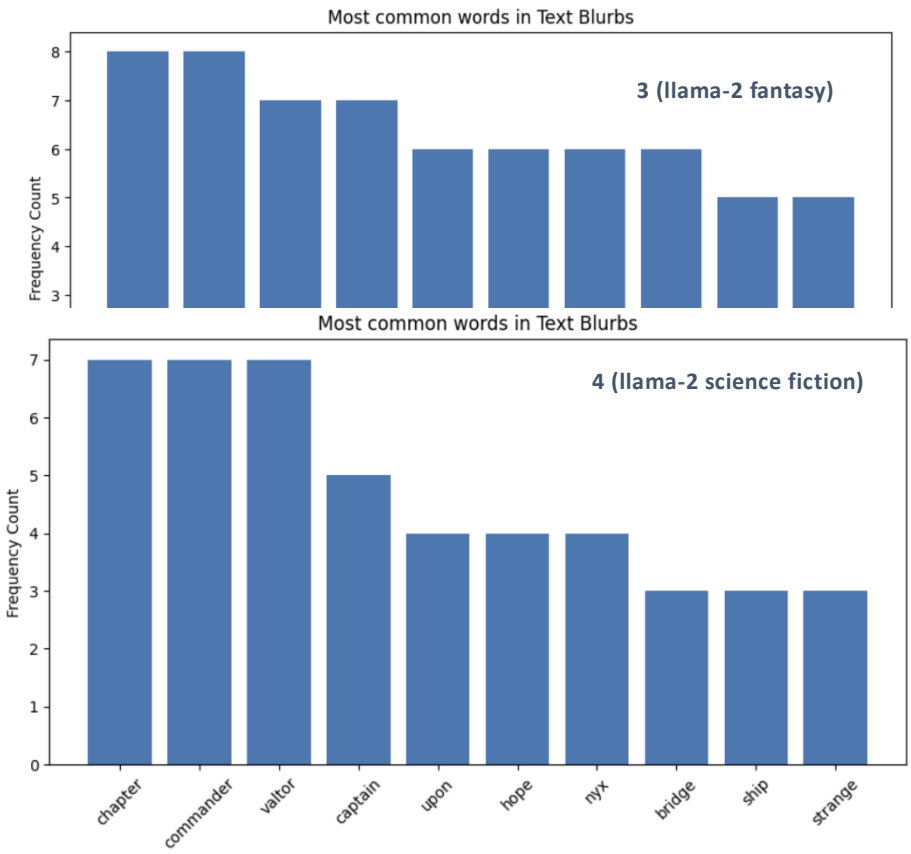
Samples of Visual Findings

I will now display some visuals for my findings. Due to the sheer size of the txt files, my computer was not able to handle large sizes of text (1500000 or higher), and I separated the processes for each file with different notebooks, identical code.

The two charts below show the differences between science fiction and fantasy in my initial text file analysis. I would assume that fantasy would use the word "upon" more often, as fantasy books are written with much older or dated vocabulary. However, I can see here, that science fiction uses "upon" much more than fantasy, telling me about the writing style of the genre. This is also largely because the files I chose from the Project Gutenberg site are quite old, and the writing style is reflected in this analysis.



I used the occurrence of the most frequently used words to generate my llama-2 stories. Here are the NLP graphs generated from these shorter stories (science fiction and fantasy). Here, I prompted the model based on my analysis, to include the most frequently used words for each genre in generating each story (frequently used words for fantasy, then for science fiction). Both genres used upon, and I did not mention the order in frequency. Still, science fiction tends to "upon" as much as fantasy. The llama-2 generated story for fantasy does use upon more, yet that is because llama-2 generated slightly more tokens than it did for science fiction. The use of my NLP findings and analysis did perform well in helping me create stories with similar results to my text files.



Academic Discussion

Now, as for academic discussion about prompting based on NLP genre analysis, I have found plenty of articles that cover NLP analysis of texts, such as finding the sentiment, summarization, and topic grouping. This article by Everton Gomede, written in 2023, discusses how using word embeddings can help generate more “emotionally captivating content” (Gomede, 2023). He also mentions that this is not without challenge, as discovering bias and ensuring diversity is imperative to ensuring better understanding of story analysis. An article by Dalton Crutchfield in 2020 provides a much more in-depth analysis into story genres using texts pulled form Project Gutenberg, yet focuses solely on that rather than applying those findings to prompting. His academic research paper provides a lot of helpful insight into genre analysis that

I most likely could have used in hindsight instead of doing my own analysis. What I gather is that while NLP analysis into genres of stories has been done many times, and has been written about many times, not much has been written about how these findings translate into prompting advice. Many prompting advice found on the internet discuss how to prompt without going into the reason behind the prompting, such as what NLP findings make these prompting tips so good.

Conclusion

To conclude, my findings here are surface level. While my analysis did prove to be helpful at generating better stories, I would like to further develop this research with a wider variety of stories to prevent bias for European literature or American literature, and to research what other individuals have done on story analysis to see how I can improve prompting stories of good caliber.

References

Crutchfield, Dalton, J. 2020. *Using Natural Language Processing to Categorize Fictional Literature in an Unsupervised Manner*. University of Denver, Digital Commons. https://digitalcommons.du.edu/cgi/viewcontent.cgi?article=2739&context=etd.

Gomede, Everton. 2023. *Word Story Embeddings: Unraveling Narratives with Natural Language Processing*. Medium. https://medium.com/@evertongomede/word-story-embeddings-unraveling-narratives-with-natural-language-processing-151827fa29f0.