Chat Regex Report

Andrei Cozma
Graduate Student
University of Tennessee
acozma@vols.utk.edu

Manan Patel
Undergraduate Student
University of Tennessee
mpatel65@vols.utk.edu

Tulsi Tailor
Undergraduate Student
University of Tennessee
ttailor@vols.utk.edu

Zac Perry
Undergraduate Student
University of Tennessee
zperry4@vols.utk.edu

I. Introduction

In this report, we examine three intriguing books: "Murder on the Links" and "The Man in the Brown Suit" by Agatha Christie, and "The Sign of the Four" by Sir Arthur Conan Doyle, all sourced from Project Gutenberg eBooks. We use a unique chatbot that addresses questions regarding investigators, crimes, perpetrators, and co-occurrences, unraveling the narratives within these literary gems. We've improved text analysis with innovative design choices in text pre-processing and a unique chatbot algorithm. Even though we faced challenges in data pre-processing and algorithm development, our dedication to error handling and testing guarantees a dependable digital reading experience. This report reflects our meticulous process and inventive solutions in exploring these captivating literary works.

II. Novel Analysis and Findings

This report will respond to the provided novels by addressing the following inquiries.

1) When does the investigator (or a pair) occur for the first time
2) When is the crime first mentioned - the type of the crime and the details
3) When is the perpetrator first mentioned
4) What are the three words that occur around the perpetrator on each mention (i.e., the three words preceding and the three words following the mention of a perpetrator)
5) When and how the detective/detectives and the perpetrators co-occur
6) When are other suspects first introduced

A. Murder on the links

"Murder on the Links" is a detective fiction novel by Agatha Christie. Hercule Poirot, the detective, is first introduced in the first chapter, "Fellow Traveller" (sentence 7). The crime itself is first mentioned in chapter three, titled "At the Villa Genevieve" (sentence 90), where Leonie, a young maid, discovers the stone-dead body of M. Renauld. Toward the novel's end, in chapter 27 (sentence 157), Hercule Poirot assembles all the suspects to reveal the solution to the murder. Additionally, our analysis reveals that there are numerous suspects involved in this murder mystery, including M. Bex (chapter 3, sentence 841), Leonie Oulard (chapter 3, sentence 939), Denise Oulard (chapter 3, sentence 984), Jack Renauld (chapter 3, sentence 988), and many more. Marthe Daubreuil is eventually revealed as the perpetrator.

B. The Man in the Brown Suit

In "The Man in the Brown Suit," Colonel Race is the investigator, introduced in the 39th sentence of the Prologue. The central crime, Nadia's stabbing, is revealed in the 127th sentence of the third chapter, with Sir Eustace Pedler as the perpetrator in the fifth sentence of the first chapter. The top three words associated with the perpetrator are "Knowledge," "Mill," and "Respecting." The detective and the perpetrator co-occur in Chapter 9, sentence 134, Chapter 24, sentence 188, and Chapter 24, sentence 84. Guy Pagett, a suspect, appears in Chapter 2, sentence 54, and Suzzane Blair, another suspect, is in Chapter 30, sentence 64.

C. The Sign of the Four

In "The Sign of the Four", by Sir Arthur Conan Doyle, the main investigator, Sherlock Holmes, is first mentioned in the first sentence of chapter one. Although there are many crimes, we discovered that the main crime in the novel is the robbery and theft of valuable jewels and treasure. The robbery is discovered to have happened in Chapter five, sentence 175. The perpetrator, Jonathan Small, is first mentioned in Chapter three, sentence 53 when Holmes is given a small note listing the perpetrator's name. When analyzing the words surrounding the occurrences of the perpetrator, we discovered that many descriptive words were used, such as "shrewd" and "wooden-legged." We also found that the detective, Holmes, and the perpetrator, Small, occur in the same sentence twice. This happens in chapters 11 and 12, sentences 9 and 36, respectively. Finally, through further analysis, we also found the other suspects' first occurrences within the novel. This includes Captain Morstan in Chapter two, sentence 31, and Major Sholto in Chapter two, sentence 43.

III. Design Choices for Implementation

A. Pre-Processing

Text pre-processing is crucial in preparing textual data for analysis and natural language processing tasks. The process begins by reading text files and proceeds through various stages to ensure the data is clean and standardized.

The first step addresses extra whitespace in the text, such as consecutive spaces and newlines. We limit consecutive

spaces and line breaks to a consistent amount and trim leading and trailing whitespace on each line. This is critical for preserving a consistent text structure and making subsequent steps easier.

Next, we remove the non-essential text that is found between the "START_OF_THE_PROJECT" and "END_OF_THE_PROJECT" strings typically found in Project Gutenberg eBooks. By removing these sections, we ensure that only the essential text of the chapters is retained for subsequent analysis.

Special tokens are inserted at the beginning of each chapter to aid in parsing the data later for analysis. This is done through regex substitution, using a pattern that matches the chapter titles. The chapter delimiter serves as a marker that precedes the title at the start of each chapter, allowing for an accurate count of chapter numbers.

In the original text files, sentences are often truncated due to line breaks, likely a result of being scanned directly from books in a column-formatted layout. To rectify this, we replace newlines that isolate characters within the same paragraph with spaces. This unifies the text into coherent paragraphs, making it easier to parse and analyze.

The novels we analyze are encoded in Unicode, including special symbols and diacritics that complicate text parsing. To address this, we normalize the text by converting it to Normalization Form D (NFD), which decomposes characters into their base forms and separates diacritical marks. Then, we strip the diacritical marks while preserving the base character. Finally, we translate various special Unicode symbols, such as curly quotes, ellipses, dashes, and others, to ASCII equivalents.

Lastly, we insert special tokens at sentence ends to assist in parsing and enable more detailed analysis. We use regex substitution, devising a pattern that aims to identify sentence endings while considering various edge cases like acronyms and honorifics. While certainly not perfect, this approach works well for our specific needs.

B. ChatBot Algorithm

This chatbot processes a given text dataset, organized into chapters and sentences, to extract information about specific terms based on predefined regular expressions. It employs the RegexPatterns enumeration to define various regex patterns that are the foundation for pattern matching and query responses.

The chatbot offers a range of capabilities to users, including special commands such as "help" and "quit." Furthermore, it can analyze the text dataset by responding to user queries. For example, it can provide information about the first mention of a term in the text, retrieve words surrounding a specific term, and identify and format co-occurrences of two terms. The chatbot also provides a friendly greeting and offers example questions to help users get started.

This chatbot is a versatile tool for text analysis, enabling users to explore and understand the text dataset more comprehensively. It leverages regular expressions and text-processing techniques to facilitate interactions and provide valuable insights into the data, making it a helpful tool for textual exploration and research.

IV. Challenges Encountered

- Data
  - Processing books from a digital library presents many challenges. For example, varying chapter heading formats and unconventional Table of Contents structures required adaptable Regex patterns to handle. When adding search terms to the text, ensuring their correctness was crucial. Preprocessing tasks, such as these, are vital for the chatbot to perform effectively and efficiently.
- Question Answer Parsing
  - Question Answer Parsing is important for two reasons. First, it needs flexible rules to understand different questions well. Second, it must correctly break down questions to give good, understandable answers. Both of these things are needed for friendly and helpful question-answering systems.
- Bot Algorithm
  - The text is preprocessed into chapters and sentences and organized for efficient lookup. Regular expressions identify user commands using unique patterns. This helps the ChatBot accurately respond to recognized commands. The ChatBot engages in a straightforward dialogue: it awaits a message, preprocesses and matches it to perform actions, and responds accordingly. It informs the user if a message needs clarification or is beyond its capabilities.
- Error handling/testing/debugging
  - In our project, error handling, testing, and debugging are pivotal for a smooth user experience. We have established a robust logging system to uncover unintended bugs and erroneous data entries. Careful insertion of search term tags and consistent formatting during pre-processing improves data quality. Our AI gracefully handles unsupported user input. Additionally, we have set up automated testing through a command-line argument to ensure a reliable and error-free product.

V. Conclusion

Overall, our in-depth analysis of the three novels has shed light on the intricacies of their narratives. Through a structured approach, we unveiled the introductions of investigators, the emergence of crimes, the first mentions of perpetrators, and the co-occurrences of detectives and wrongdoers. This exploration, complemented by our text preprocessing and innovative chatbot algorithm, has enriched our understanding of these literary masterpieces. Despite the challenges faced in data pre-processing and question-answer parsing, our commitment to quality assurance and error handling has resulted in a reliable and user-friendly digital reading experience.

REFERENCES

[1] Doyle, A. C. (2000). The Sign of the Four. Project Gutenberg. https://www.gutenberg.org/ebooks/2097

[2] Christie, A. (2019). The Murder on the Links. Project Gutenberg. https://www.gutenberg.org/ebooks/58866

[3] Christie, A. (2020). The Man in the Brown Suit. Project Gutenberg. https://www.gutenberg.org/ebooks/61168

[4] O'Reilly Media, Inc. (n.d.). Regular expressions cookbook. O'Reilly Online Learning. https://www.oreilly.com/library/view/regular-expressions-cookbook/9780596802837/ch06s09.html

[5] Bleier, S. (n.d.). NLTK's list of English stopwords. Gist. https://gist.github.com/sebleier/554280