CRIME NOVEL PLOT ANALYSIS WITH REGEX

Ori Wu, Sam Aba, Prachi Patel
Special Topics in Natural Language Processing (NLP)
Electrical Engineering and Computer Science (EECS)
The University of Tennessee
Knoxville, TN 37996
ywu70@vols.utk.edu
saba@vols.utk.edu
ppatel91@vols.utk.edu

September 27, 2021

ABSTRACT

We have conducted the analysis of various characters from five books of Maurice Leblanc. Mostly, the purpose of this analysis is to find frequencies of occurrence of protagonist and perpetrator(s) across the novel, the mention of the crime, and other circumstances surrounding the antagonist. We handled this analysis in various categories which defined in each function. We have started with dividing each book texts in chapters and then divided each chapters into sentences. After obtaining the sentences, we tokenized them by Regex and got the frequency of the words and specially extracted characters. Later, used most frequent characters to identify who they are and answer certain questions related to crime and plot.

***K*eywords** Text analysis · Word frequencies · NLP

1 Introduction

Before we begin discussing our methods of analysis, we first go into depth about the data collection and preprocessing we performed on the texts themselves. We describe the process of individually reading the books as well as using external resources [1][2] to gather relevant information. Because of our extensive use of regular expressions, we skipped many of the preprocessing steps generally associated with natural language processing. Instead we discuss the approach used to divide the chapters and sentences using complex but robust keys built from regular expressions.

Following this, we divide our analysis of the works of Maurice Leblanc [3] into six parts. The first of which identifies the first appearance of the investigators in the novel. This is followed by the in-text locations of the suspects, crimes, and perpetrator in the novels, as well as a listing of the words directly adjacent to the perpetrator.

Finally, we discuss the patterns we were able to discern from the novels and our analyses and theorize about whether these can be applied to other books by Leblanc or even in the same genre.

2 Approach

Instead of analyzing the entire book directly, at first, we prefer to divide the text of whole book into a reasonable size, namely the chapters and sentences, by using Regex grammar. The chapter-sentence division allows us to manage the text easily.

In the chapter division, one common pattern of Maurice Leblanc's books we noticed are the Roman numerals in contents. The title of a chapter is always wrote as "CHAPTER II. The Blue Diamond", "V. THE RED SILK SCARF 138" or "III. LUPIN'S WAY". Since the chapter titles have different format, beginning and ending, we proposed two methods to find the chapter titles. The first one is to use Regex pattern to search all text which begin with Roman numeral and one dot.

However, the main body may also have similar pattern, such as "Louis XIV.". In order to solve this issue, we develop the second method which search any text between two Roman numerals in order. From pair (I,II) to (XXXIX,XL), all text matched the requirement would be selected as the candidate. By combining two method, the correct chapter titles can be found and used to identify the text in one chapter. The chapter division function will split the book text into several chapter text according to chapter number.

The next step is the sentence division. Considering the long dialogue existed in the novel, one sentence can start or end with quotation marks. There are four patterns designed to identify a sentence, ("...."), ("..."), (...") and (....). Normally a sentence would be end with period mark, but we do not expect it stopped at "Mr." or "Mrs.", especially there are more prefix like "Mme." and "Mlle." in a French book. To avoid wrong division, we use other pattern like "Mme" to replace "Mme." before the sentence processing and reverse this operation after processing. The sentence divider is applied to the chapter text and return a group of sentences. Then all groups would be integrated into a list, sentence_total. For example, to visit the 86th sentence in chapter 5, sentence_total[5][85] is the required sentence. This chapter-sentence structure can quickly locate a sentence.

Since there are not existed open-source plot summary of "Lupin" series, we have to summarize the story plot by ourselves. The story in the novel revolves around the characters, and the characters are always the key component of one sentence. With the chapter-sentence division tool, we already preprocess the text into sentences and are prepared for the character analysis. The named character liked "Lupin" is a capitalized word and could be searched out by the Regex pattern like "[A-Z][a-z]". Here we create a class with attributes "Name", "Count" and "Position" to store the searched word. Each time a word has been found in a sentence, its "Count" will be plus by 1 and its "Position" will be append one array [Chapter number, Sentence number]. The searched words will be sorted by the frequency of occurrence as shown in Figure 1 according to their "Count".

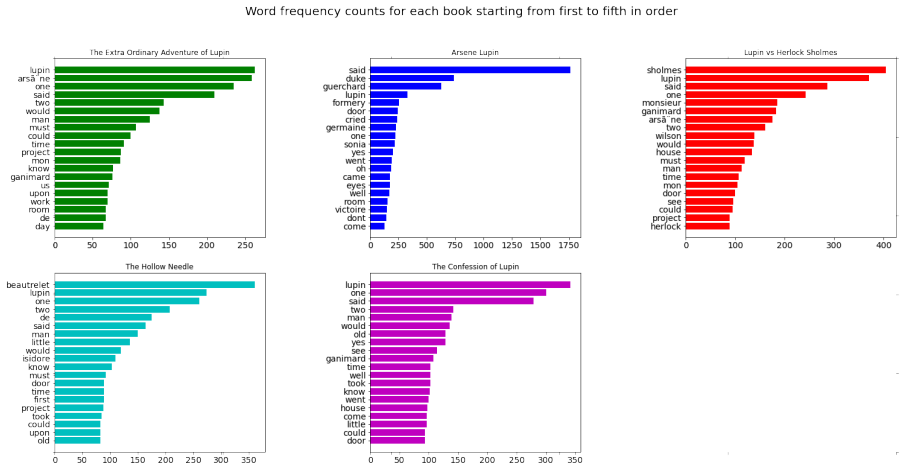

Figure 1: Word frequency counts for each book starting from first to fifth in order

After removing the normal word like "one" or "said" from the searched word, the remaining words should be characters or important objects. Next, the "Position" of the word can directly indicate which sentence contains this word. By quickly checking the sentence related to this word, we can define the detective, suspect and perpetrator in the story. Meanwhile, the crime and the plot are clearly demonstrated in front of us.

3 Results

In the Table 1 below, we list the first appearances of the detective in the story. In the case of our 5 books, the at least one investigator appeared in the first chapter 4/5 times. This occurrence happens often enough that it is safe to hypothesize that the investigator will always appear here. It is also worth noting that while Ganimard is not always the primary investigator of Lupin's crime, he still always appears in the story as a rival.

Looking at Table 2, we see that the crimes in each book vary that it there does not seem to be a large generalization to be made. Regardless, Lupin's own epithet describes him as a "gentleman thief", so it is no surprise the frequency of theft in the list of crimes. The crimes do not necessarily occur at the beginning or the end of the novel. They are much more spread out, which makes it difficult to draw conclusions from this.

Table 1: When does the investigator occur for the first time

Book Name	Investigator	Chapter No.	Sentence No.
The Extra Ordinary Adventure of Lupin	Ganimard, Folenfant, Dieuzy	1, 3	26, 341
Arsène Lupin	Guerchard, Bonavent, Dieusy	3, 10, 8	326, 274,218
Lupin vs Herlock Sholmes	Ganimard, Herlock Sholmes	1, 1	392, 165
The Hollow Needle	Beautrelet, Ganimard, Filleul	1, 2, 1	365, 66, 131
The Confession of Lupin	Ganimard, Dudouis	1, 5	379, 229

Table 2: When is the crime first mentioned

Book Name	Crime	Chapter No.	Sentence No.
The Extra Ordinary Adventure of Lupin	Theft, Attack	1, 3	127, 391
Arsène Lupin	Burglary, Plundered, Theft	8, 8, 9	181, 110, 35
Lupin vs Herlock Sholmes	Stolen, Abduction, Murder, Framed	1, 1, 2, 2	57, 296, 337
The Hollow Needle	Seize, Fight, Disguise, deceive	4, 10, 10	312, 7, 52
The Confession of Lupin	Attack	4	112

The eponymous thief in the novel is always guaranteed to commit a crime in his novel as shown in Table 3. From the table we can also gather that Lupin will appear early in the story, often as soon as the first chapter. While perpetrators of other crimes can appear in Leblanc's novels, they are few and far between. Lupin is clearly meant to be front and center.

Table 3: When is the perpetrator first mentioned

Book Name	Perpetrator	Chapter No.	Sentence No.
The Extra Ordinary Adventure of Lupin	Lupin	1	17
Arsène Lupin	Lupin, Duke	3, 1	179, 15
Lupin vs Herlock Sholmes	Lupin	1	171
The Hollow Needle	Lupin, ValmÃ©ras	2, 6	349, 27
The Confession of Lupin	Lupin	1	15

Unfortunately, there does not seem to be consistency in where Leblanc chooses to introduce the other suspects in the story. We can see in Table 4 that a suspect can be introduced as soon as chapter 1 and as late as the final chapter. Similarly, there is little information to be gained from the number of suspects or their sex.

Table 4: When are other suspects first introduced

Book Name	Suspects	Chapter No.	Sentence No.
The Extra Ordinary Adventure of Lupin	Varin, Henriette	6, 5	245, 105
Arsène Lupin	Duke, Mademoiselle	1, 2	15, 176
Lupin vs Herlock Sholmes	Clotilde, Antoinette, Bleichen, Madame, Destange	4,2,2,1,4	232,15,318,306,164
The Hollow Needle	BrÃ©doux	3	482
The Confession of Lupin	Sparmiento	10	43

The word clouds below help to visualize words that frequently occur near the perpetrator. In these novels this perpetrator is always Arsène Lupin. By qualitatively observing the graphs and ignoring the barrage of articles and helping verbs, you can see a few common words: took, laughed, and turned. Each portrays an aspect of Lupin's character; "took" because he's a thief, "turned" because he is often being chased, and "laughed" because he is overconfident having always eluded his captors.

In the following and final figure, the the appearance of perpetrators and investigators within the same sentence is counted for each novel. The x-axis shows a normalized progression through the books. Almost every line has a larger slope towards the the end of the novel. This is indicative of more interaction between the investigators and perpetrators later in the stories.

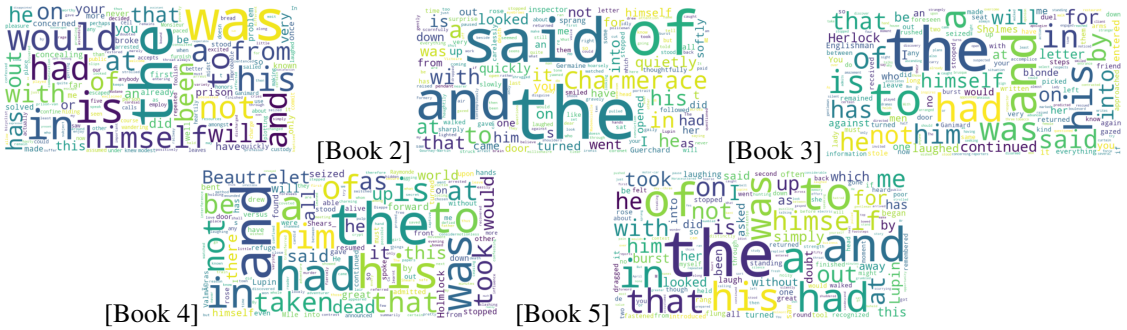

Figure 2: Word Clouds of Text Adjacent to Perpetrator (Arsène Lupin)

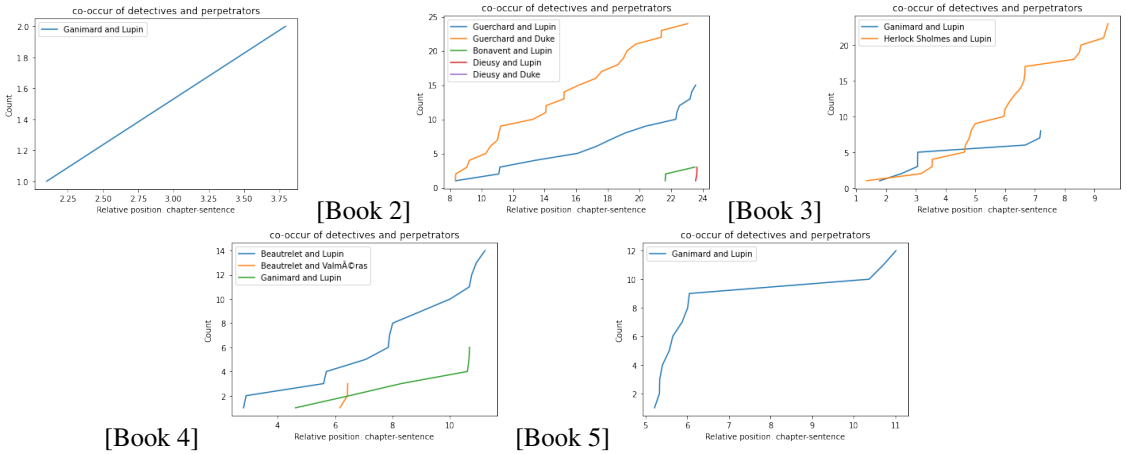

Figure 3: Normalized Co-Occurrences of Detectives and Perpetrators in the Same Sentence

4 Summary

In this study, we explore text analysis and how to perform searches for particular entities. We are able to break up texts using various functions and perform in depth analysis. We have performed in depth character analysis of each book, first determining whether characters are perpetrators, suspects, or investigators of the crime mentioned in the book. We have mainly proposed few different methods to find protagonists, perpetrators and suspects of the crime in each book and successfully found each of them. We have faced several problems during analyzing these texts such as pre-processing, dividing each chapter using regex. Since this was a complex part of the analysis as all the books in a different format, we had to apply different regex pattern for most of the books.

Secondly, once we get the answers of all crime details such as protagonists, perpetrators and suspects, we have visualised each of them and when they occurred in each book using some bar graphs, tables and world cloud which we mentioned in result section. With these we presented our analysis in a way that allowed us to identify patterns within the text. Overall, we were able to prove correlation between novels and the locations where characters and crimes are introduced within the novel. This information can be easily generalized to Leblanc's writing and can likely be used to predict the structure of his other Lupin-related works.

Our study presented some limitations that future experiment may address. As I mentioned finding pattern using regex was time consuming and difficult. Also, finding each character and role was most difficult part. Books or any text documents carry many unnecessary words and punctuation that getting the document in a readable (clean) format is very difficult. After having cleaned the files, another obstacle was to find the relationship between characters and their role within each book. Because you cannot define all possibilities of the crime in each book – murder, theft, attack, burglary, poison – it made it difficult to search for a particular pattern. We had to manually hard code search parameters; a future implementation may find a more elegant and time sensitive method of achieving the same result.

5 References

References

[1] M. Fernandez, M. Peterson, and B. Ulmer, "Extracting social network from literature to predict antagonist and protagonist," 2015.

[2] T. v. d. B. A. S. M. E. A. N. Anna Priante, Djoerd Hiemstra, "whoami in 160 characters? classifying social identities based on twitter profile descriptions." https://aclanthology.org/W16-5608.pdf, 2021.

[3] M. Leblanc, "Lupin series." http://www.gutenberg.org/, 2021.