Konstantinos Georgio, Matthew Lane, Jean Merlet
Special Topics in Natural Language Processing (NLP)
Electrical Engineering and Computer Science (EECS)
The University of Tennessee
Knoxville, TN 37996
`kgeorgio.vols.utk.edu`
`mlane42@vols.utk.edu`
`jmerlet@vols.utk.edu`
https://github.com/NLPaladins/rinehartAnalysis

September 27, 2021

ABSTRACT

Employing elementary Natural Language Processing techniques, we have explored a series of "whodunit" works by early 20^{th} century author, Mary Rinehart. The analysis of these works includes the tracking of detectives, suspects, and victims throughout the works. The analyses were conducted through the usage of regular expressions and human in the middle algorithms for a greater level of granularity concerning the works.

***K*eywords** Natural Language Processing · NLP · Regular Expressions · Regex

1 Introduction

Mary Roberts Rinehart (1876-1958) was an American author and most well known for her mystery novel writings. Rinehart first popularized the "Had I But Known" style of writing with the publication of her work "The Circular Staircase". Her prolific career, earning her the title of "The American Agatha Christie", has led to a number of narrative adaptations from stage, television, and film.

Due to the publication dates of Rineharts' works, many are hosted via Project Gutenberg, a volunteer organization dedicated to digitization, archival, and distribution of public domain cultural works.

For the analysis of Rineharts' works, the usage of Regular Expressions was employed. A regular expression (regex) exists as a series of characters that act as a particular pattern upon which to find and/or capture specifically matching text. These expressions can be as simple as verbatim words upon which to match or capture, or as complex as nested sequences identifying particular strings upon which to include or exclude. We applied various regex patterns to detect patterns in the first appearances of perpetrators, detectives, and other suspects, as well as to extract details about the crime and co-occurences of the perpetrator and detectives.

2 Approach

2.1 Repository Infrastructure

Google Colab offers a significant flexibility for multiple users to interact with the same code base in a cloud setting. Due to the distributed nature of the team dynamic, however, a more elaborate approach was taken to ensure that team members could easily branch and merge code via a GitHub repository.

Within the `books` submodule, a dedicated class was created for each book, located within `preprocessed_book.py` file. In this class, specific initialization and text parsing for each book exist, such that loaded books could act as objects containing within them all necessary data when run. Additional functions for extracting data pertaining to each book exist within `book_extractor.py`.

Logging and configuration functions exist within the `configuration` and `fancy_logger` submodules. Within the `confs` directory, the `proj_1.yml` file exists with specific data concerning each particular book.

2.2 Data Preparation and Pre-processing

In order to ensure that all users of the code base would not run into potential path commit errors within the git repository, a data loader was established to download and parse all text directly from Project Gutenberg upon each run.

Book objects were instantiated with data extracted from the `proj_1.yml`, downloaded, and parsed into raw data. The raw data were then cleaned and parsed into a list of chapters each containing a list of sentences. Two copies of the book were then stored within each object after cleaning, one in all lower case and one with unmodified capitalization.

2.3 Research

Though Mary Rinehart enjoyed popularity during her career in the early-mid 20^{th} century, 21^{st} century readers appear to be less aware of her works. No summaries, character lists, plot devices, or any data could be found concerning Rineharts' books. As each book is an approximately 6 hour read, rather than read each book to obtain the necessary background plot and character knowledge, we instead chose to employ a *human in the middle* algorithmic approach.

For each book, all unique names ranging in size from singular proper nouns to full names with titles were extracted with:

```
(?<!‘‘)(?<¡)(?<!^)[A-Z][a-z][A-Z]?[a-z][A-Z]?[a-z]+(?:(?:\s|,|.|\.\s)
[A-Z][a-z][A-Z]?[a-z][A-Z]?[a-z]+)?(?:\s[A-Z][a-z][A-Z]?[a-z][A-Z]?[a-z]+)?
(?:\s[A-Z][a-z][A-Z]?[a-z][A-Z]?[a-z]+)?
```

Note that there exist multiple conditional `[A-Z]?` regexes. This is primarily due to the existence of both Scottish and Irish names with the prefixes *Mc* and *Mac*.

An entire list of strings to exclude were then joined by the regex or operator (`|`) and applied to filter out undesirable output. Along with unique names extracted from the book, the same process additionally kept track of which names existed within each particular sentence and chapter.

Though many locations and dates were parsed out with the usage of regular expressions via the filter, we were well aware that we would likely miss a number of "names" that ought to have been extracted. With the list of unique names in hand, we then moved on to establishing the frequency of each particular name by matching on the additional data structure returned from the previous subsection (i.e. `character_progression`). An arbitrary cutoff of 10 distinct mentions removed the vast majority of names with low frequency hits.

In order to ensure that unique characters did not appear multiple times, an aliasing measure was taken with the remaining high frequency names. As many of the aliases for a particular character did not overlap (e.g. Halsey and Mr. Innes), a rudimentary form of a suffix tree was implemented on first and surnames, with a distinction on titles. It was, however, not always the case that all characters were aliased properly, and thus human interaction was required. For example, in *The Circular Staircase*, a character *John Bailey* exists, and (by sheer happenstance) the authors came across a co-occurring sentence in which it was noted that John and Jack Bailey were in fact the same person. All human assisted output was then saved to the `proj_1.yml` file, which was used as input for all of the regex analysis functions.

3 Results

The following tables denote the first appearances of the investigators, perpetrators, suspects, and the first appearance of the crimes themselves. The subsequent section includes illustrations for the timelines concerning each book. The N+ words existing around the perpetrators name are best viewed in the accompanying notebook.

Table 1: Investigators' First Appearance

Book Name	Investigator	Chapter No.	Sentence No.
The Circular Staircase	Mr. Jamieson	1	15
The Man in the Lower Ten	Richey McKnight	1	1
The After House	McWhirter	1	41
The Window at the White Cat	Al Hunter	1	251
The Bat	Anderson	1	141

Table 2: Perpetrator First Appearance

Book Name	Perpetrator	Chapter No.	Sentence No.
The Circular Staircase	Anne Watson	2	135
The Man in the Lower Ten	Mrs. Curtis	9	63
The After House	Charlie Jones	3	130
The Window at the White Cat	Ellen Butler	14	9
The Bat	The Bat	1	23

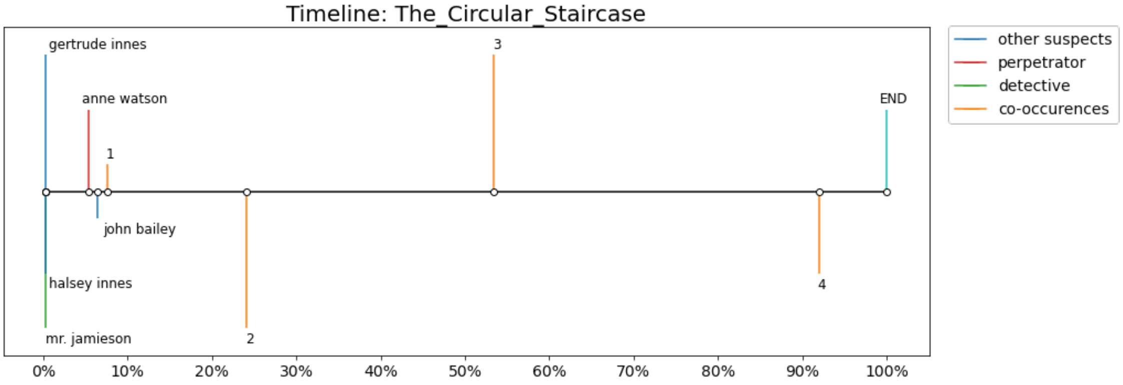
Table 3: Suspect Introduction

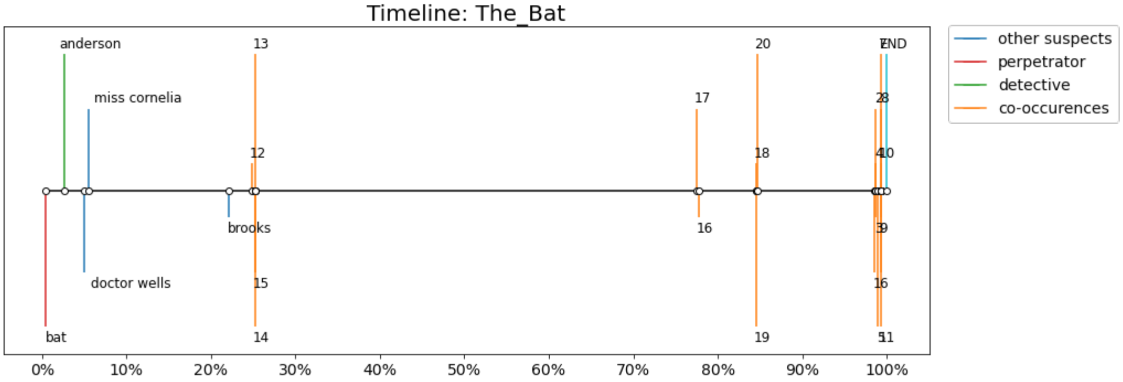
Book Name	Suspects	Chapter No.	Sentence No.
The Circular Staircase	Gertrude Innes	1	15
	John Bailey	3	47
	Halsey Innes	1	15
The Man in the Lower Ten	Wilson Budd Hotchkiss	13	49
	Heny Pickney Sullivan	5	35
	Allison West	1	162
The After House	Marshall Turner	1	100
	Mate Singleton	1	100
The Window at the White Cat	Henry Schwartz	1	241
	Harry Wardrop	3	102
The Bat	Brooks	5	140
	Miss Cornelia	2	58
	Doctor Wells	2	33

Table 4: First Crime Mention

Book Name	Crime	Chapter No.	Sentence No.
The Circular Staircase	Shooting	2	68
The Man in the Lower Ten	Stabbing	12	113
The After House	Axed	4	109
The Window at the White Cat	Shooting	7	154
The Bat	Shooting	8	213

3.1 Comparative First Appearances





In all five of the crime novels which we investigated, Rinehart consistently introduces all of her characters in the first third of the book, with a notable exception the introduction of the perpetrator in "The Window at the White Cat", which does not happen until halfway through the book. Interestingly, the books also seem to share some consistencies in the co-occurence of perpetrators and detectives, which can fairly clearly be seen in the two timelines above for "The Circular Staircase" and "The Bat". Co-occurences (defined as the matching of a detective and the perpetrator within 2 sentences of each other) tend to appear in the beginning, middle, and end of books, with large gaps in between.

As a note, the remainder of the plot summaries can be seen in the accompanying notebook.

4 Summary

While not entirely consistent, Mary Roberts Rinehart tends to introduce most of the important characters near the beginning of her works. Curiously, many of the works are in first person, and those first person narratives often are the viewpoint of an unofficial detective, with professional detectives also sometimes appearing in the rest of the narrative. One curious trend that was noticed, however, was that many of Rineharts' works have a final surmising chapter in which the dots are all connected. This comports with Rineharts' renowned "Had I But Known" style.