Visual Summarization of Stories

Jiang Zhiheng, Tan Yi Kai, Huang Junwei

Hwa Chong Institution

Teacher-mentor: Dr Chia Kok Pin, Mr Tan Choo Kee

Institutional Mentor: Associate Professor Cheong Siew Ann

Visual Summarization of Stories

Jiang Zhiheng[1], Tan Yi Kai[1], Huang Junwei[1], Cheong Siew Ann[2]

[1]Hwa Chong Institution, 661 Bukit Timah Road, Singapore 269734

[2]Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371

Abstract

A picture is worth a thousand words. By converting a story into a complex network, we produce an effective summary of a story. This visual form of the story also allows us to perform many types of analysis we cannot imagine doing with the text. In this project, we explore how to read the text of a novel into a computer using the programming language Python, identify key characters, construct the network representation of the story, and thereafter perform network-based analyses to better understand the story. Certain complex stories can have simple graphical representations, allowing them to be effectively summarised. Such graphical representations involve the use of the Python programming language to read the story and exporting data to the third-party graphing software Gephi to produce a graph.

Introduction

When we read a book, we may find it extremely time-consuming to digest every single detail of the book and develop a complete idea of the plot and the story. There are many factors contributing to the decline of reading, including a faster pace of life, a shorter attention span and others. According to a 2014 Time magazine[1] article, Americans aged 40 and above spend on average 19 minutes per day reading but those aged 18 to 25 spend just 4 minutes per day. Nonetheless, it is clear the more of the plot a reader understands, the more he will enjoy the book. Thus, we want to allow readers to enjoy the book more by helping them to quickly understand the plot by representing it through a graph, more specifically, a complex network. This will allow the reader to easily understand the gist of the story, saving them valuable time and effort. About 407,000 books are published each year, adding up to more than 8 million books over the past 20 years. Even if every person were to read a book every day, he would not be able to read all 8 million or so books. More importantly, while ordinary people are not required to read books, literature scholars have to. They will need help to decide which books are worth reading as most books are simply a rehash of plots that have been repeated infinitely many times. Additionally, this can be used for news analysis, where the key information is unlikely to change much from a piece of news to the next. For example, no matter how a reporter spins a murder case, it will always revolve around the murder, making it suitable for analysis.

Literature Review

Natural Language Processing[1] (NLP) and other forms of computational linguistics are like engineering and science. The first solves problems involving natural language analysis while

the second analyses natural language to understand why their grammar, word etymology and in general, why they are the way they are.

In computational linguistics, one of the main goals is text summarization, where the programme reads in from a raw text file to produce a summary, be it in a visual form such as a graph or a semantic web, or in text form comprising of a set of keywords. NLP is especially helpful in this aspect. It is the process whereby a machine reads natural language, in this case, English. There are many levels of difficulty for NLP, in the higher levels, the NLP is used to understand text rather than lower level NLPs such as translation. NLP can be used for information retrieval, information extraction, question-answering, summarization (higher-level NLP) and machine translation. NLP is a relatively new area of research and has much potential in the future, for machine to understand human language. As this technology is rather new, our project focuses on summarization and visualisation using NLP to help readers better understand long text by providing a short and concise summarization of the plot of the story. However, there are many limitations of Natural Language Processing [2](NLP), such as limited understanding and certain amounts of inaccuracy. NLP has limited research with regards to network summarization of text. NLP can also be used for human-computer interaction, translation, NLP interface, help systems and information retrieval. Our project would be trying to overcome these limitations to allow NLP to visualise and summarise long text.

Data

Title of Book	A Study in Scarlet	A Mysterious Affair at Styles	The Circular Staircase
Author	Arthur Conan Doyle	Agatha Christie	Mary Roberts Rinehart
First published	1887	1920	1908
First publisher	Ward Lock and Co.	John Lane	Bobbs-Merrill
Source	https://www.gutenberg.org/ebooks/244	https://www.gutenberg.org/ebooks/863	https://www.gutenberg.org/ebooks/434
No. of Words	43,414	53,507	56,378
Cleanup	Remove headers		
Format	.txt		

Methods

Our core-periphery network is a graph displaying the connections between nodes which are representing character of a story. The core of the graph should be densely connected and many of the main characters can be found here. The periphery or the outer fringes of the graph are sparsely connected and less significant characters are found here.

The development of our core-periphery network involves the following steps:

1. We have employed the use of the Natural Language Tool Kit [3] (NLTK). NLTK is a branch of Natural Language Processing (NLP) and allows us to separate individual sentences into subjects, verbs and objects with around good accuracy. It is also compatible with Python, allowing us to use it with little hassle. Identifying important and moderately important characters requires four steps:

a. Identifying proper nouns and names. This is relatively simple as proper nouns and names begin with a capital letter, making them easily identifiable with NLTK.

b. Differentiating between agents and patients. Agents are defined as characters that actively do things and drive the storyline forwards. Patients are defined as characters that are passive and wait around for things to happen. Major characters tend to be agents while minor characters tend to be patients. Agents are usually the subject of a sentence while patients are usually the object of a sentence. Thus, we use NLTK to separate the sentence into the subject, verb and object to identify the agents and patients.

c. Counting the number of times a character acts as a agent or a patient. This is done by counting the number of times NLTK identifies it as a subject or object of a sentence, which agents corresponding to subjects and patients corresponding to objects.

d. Determining the most important agents. This is done by calculating an agent-patient ratio. A higher agent-patient ratio indicates a more important agent. However, the frequency of appearance of a character is also taken into consideration as minor characters that only appear once or twice but happen to have high agent-patient ratios should not be considered as major agents.

2. Manual disambiguation of the text. There are 2 areas in which we will have to do this:

   a. We must disambiguate all the pronouns that appear in the text (e.g. he, she, it) and replace them with the name of the character they are referring to. This can be done by scanning the previous sentences for the most recent character mentioned and thus replacing the pronoun in question with the identified character.

   b. Sometimes, due to natural flaws in NLTK or different people addressing the same person in different ways, different subjects identified may actually refer to the same character. For example, NLTK may identify Sherlock and Holmes as different characters but they are referring to the same character Sherlock Holmes. Disambiguation in this case can be done manually, as such instances are relatively rare and simple to resolve.

   After that is done, their agent-patient counts are merged and their agent-patient ratios re-calculated.

3. Standardisation of the text. All names that refer to the same character in the text will be replaced by one standard name. This is to make future analysis simpler avoid confusion by NLTK.

4. Construction of the weighted, undirected network. A preliminary node and edges graph is plotted using networkx. The thickness of the edge between two characters is determined by their agent-patient ratios relative to each other or the number of times they are identified as subject or object in the same sentence by NLTK. However,

which character has a stronger influence on the other cannot be determined from the graph as it undirected.

5. Construction of the weighted, directed network. We will export our data to an external graphing software Gephi. It helps plot a weighted network, similar to networkx, but also with directions, represented by arrows, allowing us to tell between two characters, which one is the agent and which one is the patient.
6. Construction of the weighted, directed network, We used NLTK Vader Sentiment Analysis to analyse and colour the edges according to their respective RGB values, with red being the most negative sentiment, blue being the most positive, and green being neutral.
7. Having obtained the RGB values of connections between 2 characters, we plotted a graph using graphing software Gephi. We export the node and edges values through a gexf file to Gephi and it helps us plot a basic nodes and edges graph. Similar to our networkx graph, a thick edge between 2 nodes would mean a strong connection while a thin edge between 2 nodes would mean a weak connection. However, the connections now have directions, which are represented by arrows pointing at the character which is receiving the action.
8. We interpreted the plot of the story through the Gephi graphs of the 3 stories. Refer to results.

Results & Discussion

Below are the network summaries of the three stories we have summarised. All three graphs are generated using the graphing software Gephi after the steps mentioned above.

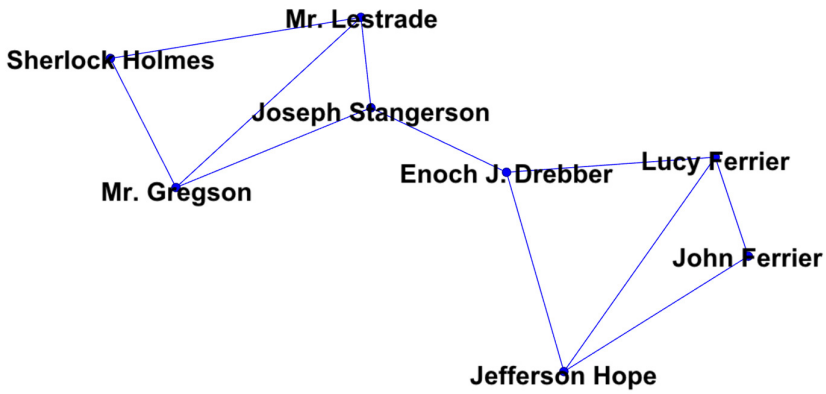

Figure 1. The undirected network summary of *A Study in Scarlet* by Arthur Conan Doyle, first published by Ward Lock and Co. in 1887.

We first created the undirected network summary, which showed a network of the characters, showing which characters had relations, but lacked the details for much analysis.

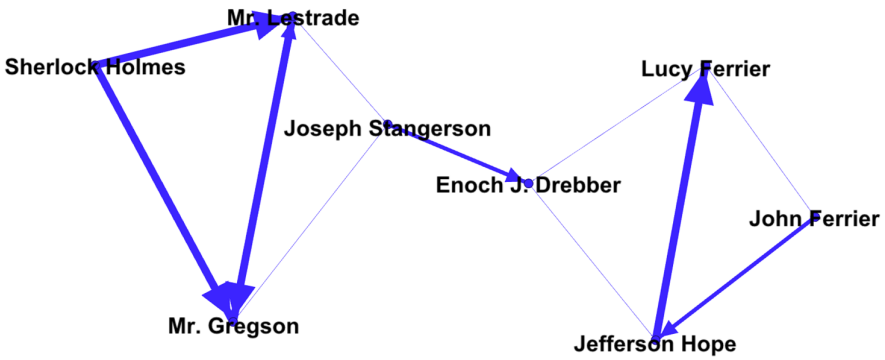

Figure 2. The directed weighted network summary of *A Study in Scarlet* by Arthur Conan Doyle, first published by Ward Lock and Co. in 1887.

We then created the weighted directed network summary of *A Study in Scarlet*. It is the simplest to understand out of all the three stories we have attempted to summarise using the network method. It comprises of two strongly connected network neighbourhoods, one involving Sherlock Holmes, Inspectors Gregson and Lestrade and another one involving Joseph Stangerson, Jefferson Hope, John Ferrier and Lucy Ferrier. The only link between the two network neighbourhoods is through Enoch Drebber, thus allowing information flow between the neighbourhoods.

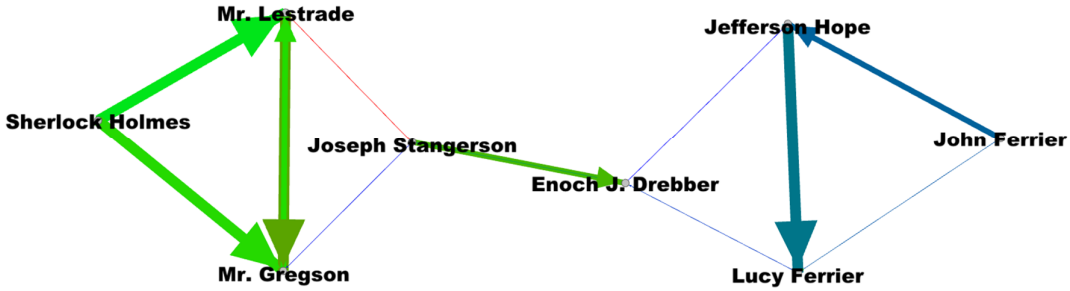

Figure 3. The directed network summary of *A Study in Scarlet* by Arthur Conan Doyle, first published by Ward Lock and Co. in 1887.

The information flow between the neighbourhoods is further verified as the link between them is neutral, as seen in the coloured network summary. It is through Drebber and the only person linked to Drebber, Stangerson that the Inspectors and Holmes find the murderer Jefferson Hope and piece together the full picture of the crime.

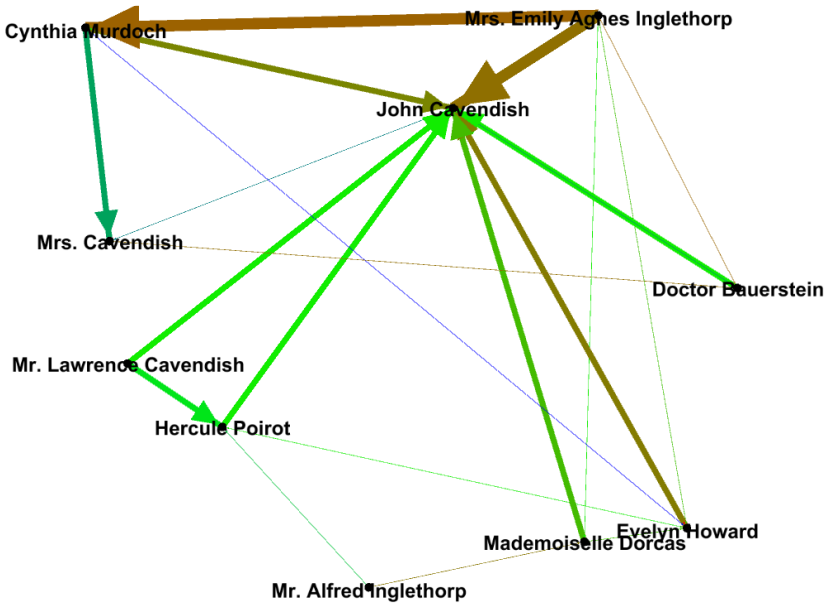

Figure 4. The directed weighted coloured network summary of *The Mysterious Affair at Styles* by Agatha Christie, first published by John Lane in 1920.

The network summary for *The Mysterious Affairs at Styles* is considerably more difficult to interpret than for *A Study in Scarlet*, as it involves more characters and more network neighbourhoods. Although the central character, John Cavendish, can be easily determined, the overall plot of the story is difficult to understand. There are multiple network neighbourhoods involving three or four characters, each featuring a part of the overall story. It seems that the interactions between the network neighbourhoods is largely limited to via the central character John Cavendish. How the detective, Hercule Poirot, finds the murderer, Alfred Inglethorp is unclear although weak connections do exist.

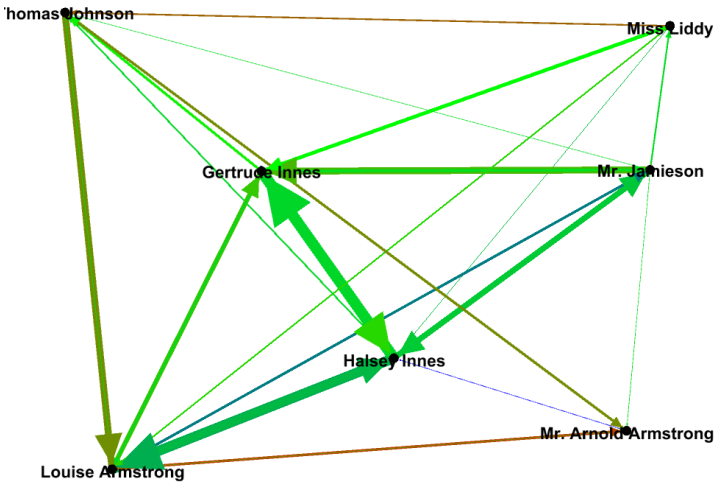

Figure 5. The directed weighted coloured network summary of *The Circular Staircase* by Robert Rhinehart, first published by Bobbs-Merrill in 1908.

The network summary for *The Circular Staircase* is the most confusing of the three books that we have summarised. Not only are the individual network neighbourhoods not clearly defined, there is no clear central character. This may be due to two reasons:

1. There is a case of joint agency, which means that two characters execute the same actions together and are referred to as if they were a single entity very often. However, in the network summary, they are represented as if they are separate characters, even though their network connections are largely similar.
2. The narrator plays a significant role in the story but not all of the narrator's actions are significant. We have thus decided to omit the narrator from the network summary. If we included the narrator, it would have been mostly an island node with little connections with the other characters.
3. The story itself is different from the other two stories. In the first two stories, the murder of a character and the subsequent investigation is the focus. However, in the third story, the murder is only the surface. It is part of a wider intricately planned plot to scam money out of the villagers and the focus is shifted to figuring out who is involved in this plot. Additionally, there is not one or two murders in the previous two novels but six, making the story all the more complicated.
4. There exist subplots of love stories to form the bigger story, making the interactions less focused on the detective and crime aspect of the story.

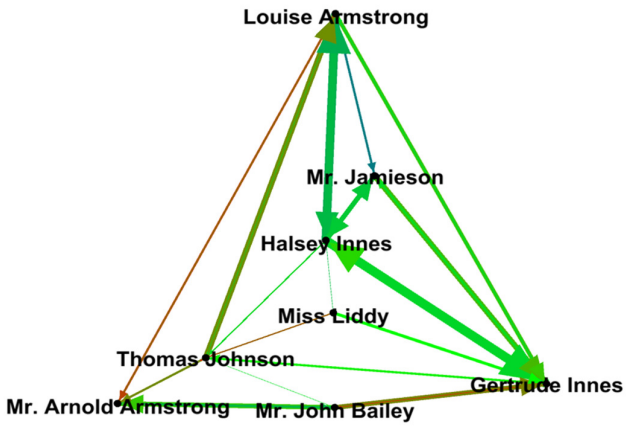

Figure 6. The directed weighted coloured planar network summary of *The Circular Staircase* by Robert Rhinehart, first published by Bobbs-Merrill in 1908.

Figure 6 shows a planar representation of the graph of *The Circular Staircase*. We computed the Planar Maximally Filtered Graph (PMFG) for the network in order to filter out less important edges and focus on the more important edges to analyse. We can observe a four-character network neighborhood between Louise Armstrong, Mr Jamieson, Halsey Innes and Gertude Innes more clearly from the planar graph which shows important interactions to make these characters strongly connected.

Error Analysis

We find that there are certain areas that we could have improved on during our research

1. Missing out characters. We may miss out characters' interactions, especially those between characters that appear infrequently. One way to resolve this is to programme a Python script and pull out all sentences with some characters A and B and manually count their interactions
2. Disambiguation difficulties. Two different names can refer to the same character but we have no way of knowing this at the start without reading the book and thus must treat them as separate characters. This allows the same character to have two different agencies, which is confusing for plotting the agency-patiency ratio graph later on.

Conclusion

Overall, our project has been largely a success, given the severe technical limitations that we faced. We managed to successfully create relationship graphs first using the networkx and then in Gephi for all three detective novels that we analysed. The relationship webs displayed in the graphs allowed us to find the main characters of the story, whether they had a positive or negative relationship as well as how strongly connected they were. However, due to technical limitations, we could only clearly determine the plot by looking at the graph for one of the three stories, namely A Study in Scarlet. The other two graphs were simply overly complicated and contained many complex network neighbourhoods. Other factors such as the writer's style, ineffective sentiment analysis and characters with multiple aliases or joint agency also affected the result and are future areas of development.

Bibliography

1. Stephen, B. (2014, June 22). Americans Read 19 Minutes per Day. Retrieved from http://time.com/2909743/americans-reading/
2. (Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc, https://surface.syr.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com.sg/&httpsredir=1&article=1019&context=cnlp)
3. (Gerhard, F. (1996, November 11). Natural Language Processing. Retrieved February 14, 2018, from http://l3d.cs.colorado.edu/courses/AI-96/nlprocessing.pdf)
4. (Lin, J. (2008). Scalable language processing algorithms for the masses. Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP 08. doi:10.3115/1613715.1613769)
5. (Project Gutenberg) http://www.gutenberg.org/
6. Bird, S., Klein, E., & Loper, E. (n.d.). Natural Language Processing with Python. Retrieved May 13, 2018, from http://www.nltk.org/book/

Acknowledgements

We would like to express our sincere appreciation towards Dr Chia Kok Pin and Mr Tan Choo Kee for overseeing our project. Most of all, we would like to show our most heartfelt gratitude towards our mentor Associate Professor Cheong Siew Ann for his unwavering support and guidance throughout this project. Last but not least, we would like to thank all of our classmates and friends who offered help throughout the entire course of this project.