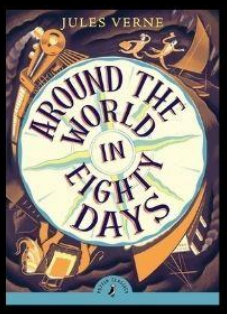
80 days around the world

Social Media WorkShop



Shlomi Shor III 316328236

Final Work – Social Media WorkShop

https://colab.research.google.com/drive/1bhE5WkiOO_u6gTWRgwnVBM2k1bWmgyob#scrollTo=KYvA2ulqimrI

Description -

1. **Introduction:**
   a. My project constructs social networks from books.
   The user shares a book they are interested in as a PDF file, and the machine reads the book, identifies the characters, and builds the social network of the book for them. If two characters appear within a gap of 3 lines between them, a connection is established between the characters. The more frequently they appear within this gap, the stronger the connection between them becomes.
   b. It is important to note that the machine sometimes creates duplicates of characters, as sometimes the same character is referred to by different names. In the book I based my project on, "Around the World in 80 Days," the main character Phileas Fogg is also referred to as "Mr. Fogg," "Fogg," "Phil," and more. Therefore, the machine does not process this perfectly and creates duplicates, which I had to merge, as you will see in the second part of the project.
   c. In this work, I aim to showcase the capabilities of my machine without any adjustments, which you will see in the first part. In the second part, I analyze the book based on the actual characters.

2. **Functions:**
   a. **Function 1 ("normalize_name") –** Removes newlines and extra spaces from a given name.
   b. **Function 2 ("extract_character_names") -** Extracts all named entities labeled as "PERSON" from a PDF using spaCy and returns them along with the full text.
   c. **Function 3 ("group_similar_names")** - Groups similar names based on a Levenshtein similarity threshold, merging very similar groups into a single canonical name.
   d. **Function 4 ("extract_and_replace_names")** - Replaces aliases in the text of a PDF with their canonical names and extracts lines where character names appear.
   e. **Function 5 ("extract_and_locate_names") -** Locates and records the appearances of character names in a PDF by page and line number.
   f. **Function 6 ("connection_strength1") -** Calculates the connection strength between two characters based on the ratio of their common appearances to their total appearances.
   g. **Function 7 ("connection_strength2") -** Computes the connection strength between two characters based on the number of times they appear within three lines of each other on the same page, relative to their total appearances.

h. **Function 8 ("calculate_connections") -** Calculates the connection strength between all pairs of characters in the given dictionary using connection_strength2 and returns the results.

Part I- Idea

1. **Question 1 -**

   Proposed project questions –

   a. **Who are the main characters in the book?**
   Identifying the main characters of the book is fundamental to understanding the key actors and key points within the book. Central characters influence the plot, on other characters and on their relationships.

   b. **What are the social circles of the characters?**
   Analyzing the social circles of characters helps in understanding the relationships and networks within the book. By examining these circles, I can see how characters are grouped, who belongs to which cluster, and how these groups interact with each other, revealing the social structure and dynamics in the book.

   c. **What is the strength of the relationship between the characters?**
   Assessing the strength of relationships between characters allows for a deeper understanding of the connections and bonds that exist in the story. In social network analysis, this is akin to measuring the weight or strength of ties between nodes. Strong relationships may indicate frequent interactions or significant influence, which can be critical for understanding the flow of information or influence within the network.

Part II- Theory

2. **Question 2-**

   a. **Who are the main characters in the book?**
   Centrality Measures - Identifying main characters is akin to finding central nodes in a network. Centrality measures help us understand which nodes (characters) have the most connections, act as key bridges, or are closest to all other nodes.
   Contribution: These concepts help us identify which characters are most influential or connected within the narrative, akin to identifying influential nodes in a social network.

   b. **What are the social circles of the characters?**
   Communities and Clustering - Social circles are like clusters or communities in a network. Clustering algorithms can detect groups of nodes that are more densely connected to each other than to the rest of the network.
   Contribution: Understanding social circles provides insights into the sub-structures within the network, revealing close-knit groups or factions within the narrative.

c. **What is the strength of the relationship between the characters?**

Edge Weights and Strength - Relation: The strength of relationships between characters can be represented as edge weights in a network, reflecting the frequency and intensity of interactions.

Contribution: This helps quantify the strength of connections and can highlight the most significant relationships, much like measuring tie strength in social networks.

Part III- Data

3. **Question 3:**

The dataset includes the text from the book 'Around the World in Eighty Days' by Jules Verne. This dataset is relevant because it contains the narrative in which the characters interact, allowing us to map these interactions and analyze the social network within the story. I used the PDF version of the book as my primary source, which was processed using several Python libraries, including PyMuPDF for text extraction and spaCy for named entity recognition (NER). Using the fifth function I extracted the positions of each character by page and line number. Finally, I defined a relationship between characters based on their appearance within three lines of each other on the same page and calculated the strength of the relationship based on the frequency of co-occurrences. The limitation of the initial data set was that the software did not know how to merge 100% between all the nicknames of the characters in the book. For example - the main character Phileas Fogg has several names like - Mr. Fogg, Fogg, etc., the software needed manual intervention and was unable to merge all the nicknames of a certain character.

A. **The process of obtaining the data set –**

The book -

https://drive.google.com/file/d/1aDaaNmaBObGomreaQL0RVGRicNLa-i0v/view?usp=drive_link

```
[3] #Loading the book

    file_id = '1m3zgJ2877lGePrFRqYZJ4kGEwaEKTJFj'
    url = f'https://drive.google.com/uc?id={file_id}'

    pdf_path = 'my_book.pdf'
    gdown.download(url, pdf_path, quiet=False)
```

```
Downloading...
From: https://drive.google.com/uc?id=1m3zgJ2877lGePrFRqYZJ4kGEwaEKTJFj
To: /content/my_book.pdf
100%|██████████| 1.32M/1.32M [00:00<00:00, 11.7MB/s]
'my_book.pdf'
```

```
[5] #Book Information:

    df_book = fitz.open(pdf_path)
    print('Title:', df_book.metadata['title'])
    print('Author:', df_book.metadata['author'])
    print('Pages:', df_book.page_count)
```

```
Title: Around the World in Eighty Days
Author: Jules Verne
Pages: 242
```

Limitations:

```
[6] # Union between similar names:

    nlp = spacy.load("en_core_web_sm")
    pdf_path = 'my_book.pdf'
    character_names, full_text = extract_character_names(pdf_path)
    grouped_names = group_similar_names(character_names)

    # Print the grouped names
    print("Grouped Character Names:")
    for key, names in grouped_names.items():
        print(f"{key}: {names}")
```

```
Grouped Character Names:
Saint Joseph: ['Saint Joseph']
Bordeaux: ['Bordeaux']
Dutchmen: ['Dutchmen']
Medicine Creek: ['Medicine Creek']
Juggernaut: ['Juggernaut']
Julesburg: ['Julesburg']
Burdivan: ['Burdivan', 'Burdwan']
Nassik: ['Nassik']
Thomas C. Durant: ['Thomas C. Durant']
José Menéndez: ['José Menéndez']
Samuel Fallentin: ['Samuel Fallentin']
Stuart: ['Stuart']
Elder Hitch: ['Elder Hitch']
Joe Smith: ['Joe Smith', 'Joseph Smith']
Jules Verne: ['Jules Verne']
Aureng-Zeb: ['Aureng-Zeb']
Mew: ['Mew']
Fogg: ['Fogg']
Sombreros: ['Sombreros']
Speedy: ['Speedy']
Mandiboy: ['Mandiboy']
Japanesed: ['Japanesed']
Aouda: ['Aouda']
Pereire: ['Pereire']
Suez: ['Suez']
Monsieur Fogg: ['Monsieur Fogg', 'Monsieur Fix']
Calais: ['Calais']
Kent: ['Kent']
Parbleu: ['Parbleu']
George Towle's: ['George Towle's']
XXIX: ['XXIX']
James Strand: ['James Strand', 'James Strand's']
Francis: ['Francis']
Queen: ['Queen']
Andrew Stuart,—"he: ['Andrew Stuart,—"he', 'Andrew Stuart']
Phileas Fogg: ['Phileas Fogg', 'that Phileas Fogg', 'Phileas Fogg's']
```

An example of a limitation - the software recognizes names of places such as the Suez

An example of a limitation - the Tanuka does not unite all the names 100%

Identify the character in the book by page and line:

```
[7] # Identify the character in the book by page and line:

    results, modified_text = extract_and_replace_names(pdf_path, grouped_names)
    for result in results:
        print(f"Character: '{result['character']}' (Alias: '{result['alias']}') found on page {result['page']}, line {result['line']}: {result['text']}")

    with open("modified_text.txt", "w") as file:
        file.write(modified_text)

    print("Character names unified and modified text saved.")
```

```
Character: 'Jules Verne' (Alias: 'Jules Verne') found on page 1, line 6: JULES VERNE
Character: 'Verne' (Alias: 'Verne') found on page 1, line 6: JULES VERNE
Character: 'George Makepeace Towle' (Alias: 'George Makepeace Towle') found on page 1, line 9: GEORGE MAKEPEACE TOWLE
Character: 'Towle' (Alias: 'Towle') found on page 1, line 9: GEORGE MAKEPEACE TOWLE
Character: 'Jules Verne' (Alias: 'Jules Verne') found on page 2, line 8: Author: Jules Verne, 1828–1905
Character: 'Verne' (Alias: 'Verne') found on page 2, line 8: Author: Jules Verne, 1828–1905
```

Unification of all the characters according to the relevant nicknames for each character, carried out with the help of ChatGPT:

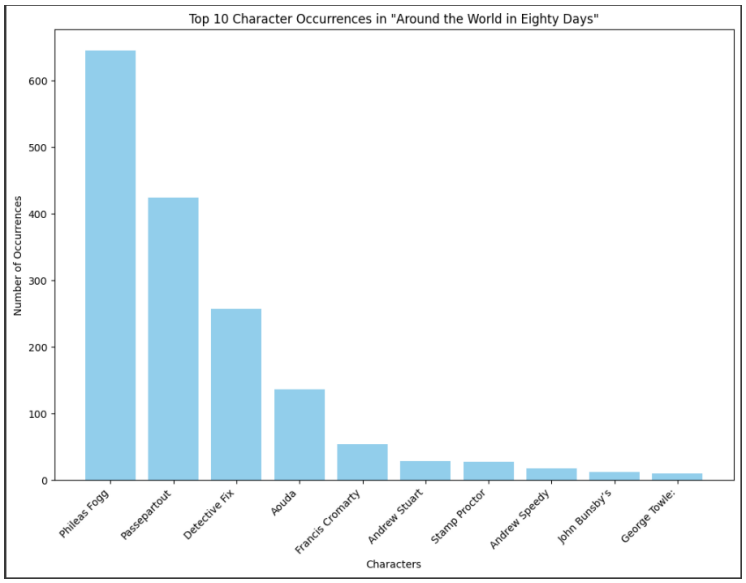
```
[20] # Unification between the names of the characters, carried out with the help of ChatGPT:

     real_characters_location = {
                                'Phileas Fogg': [(3, 6), (3, 11), (3, 12), (3, 23), (3, 26), (3, 28), (4, 6), (4, 15), (4, 19),
                                'Passepartout': [(3, 6), (3, 8), (3, 12), (3, 20), (3, 24), (4, 4), (4, 15), (4, 17), (4, 22),
                                'George Towle:': [(1, 9), (2, 9), (2, 12), (62, 34), (104, 35), (111, 35), (122, 36), (135, 34)
                                'James Forster': [(10, 14), (10, 28), (11, 40), (185, 24), (185, 33)],
                                'Andrew Stuart': [(17, 7), (17, 15), (17, 21), (19, 8), (19, 11), (19, 15), (19, 24), (19, 33),
                                'John Sullivan': [(17, 8), (20, 6), (22, 3), (22, 23), (236, 25), (237, 21), (238, 21)],
                                'Francis Cromarty': [(58, 17), (58, 19), (58, 32), (59, 9), (59, 23), (59, 24), (59, 28), (59,
                                'Aouda': [(76, 8), (83, 5), (84, 7), (84, 31), (85, 31), (86, 11), (86, 16), (87, 3), (89, 16),
                                'William Hitch': [(172, 31), (173, 13), (173, 29), (174, 9), (175, 16)],
                                'Joseph Smith': [(173, 15), (173, 34), (174, 3), (174, 24), (175, 9), (208, 20)],
                                'Thomas Flanagan': [(17, 9), (17, 13), (19, 23), (236, 27), (237, 4), (237, 17), (238, 14)],
                                'Detective Fix': [(3, 16), (4, 11), (4, 15), (4, 19), (5, 15), (30, 35), (32, 7), (33, 7), (33,
```

Part IV- Methodology

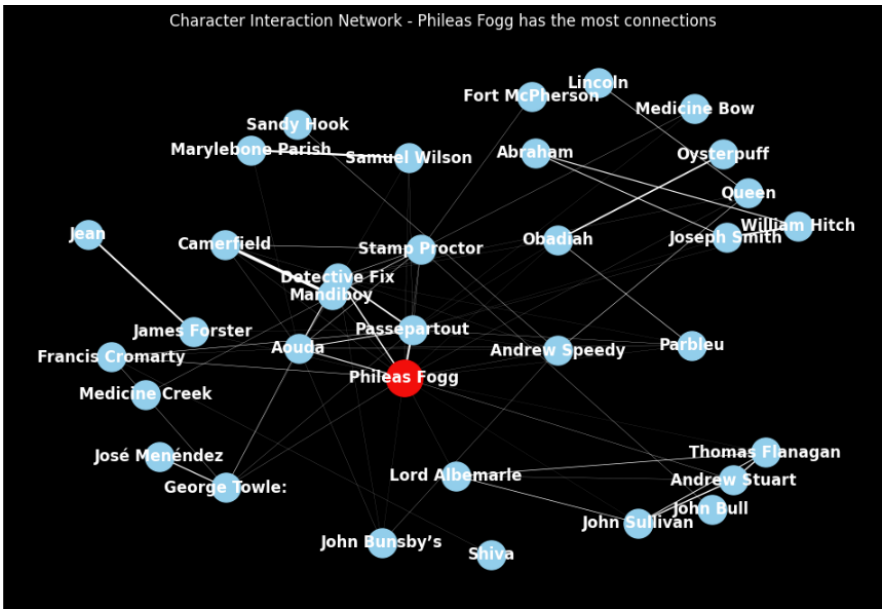
4. **Question 4:**

   A. **First Question - Who are the main characters in the book?**

   1) **Frequency of Appearance:** I counted the number of times each character appears in the book. The character who appears the most frequently is likely the main character. This is based on the assumption that main characters are more integral to the storyline and thus mentioned more often.

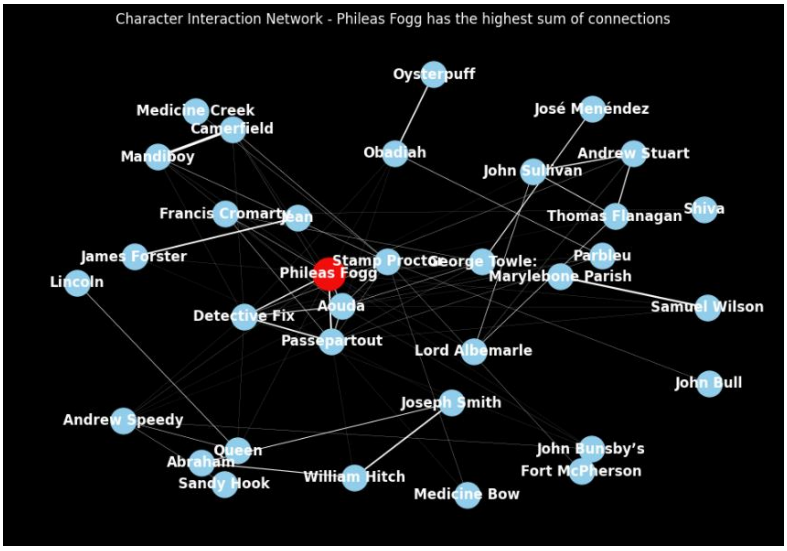


   2) **Network Connectivity:** I analyzed the social network of characters to see which character is linked to the most other characters. The most connected character in the network is Phileas Fogg. Here, I used the theory of social network analysis, specifically the concept of degree centrality, which measures the number of direct connections a node (character) has. A higher degree centrality indicates a character's prominence and potential influence in the narrative.



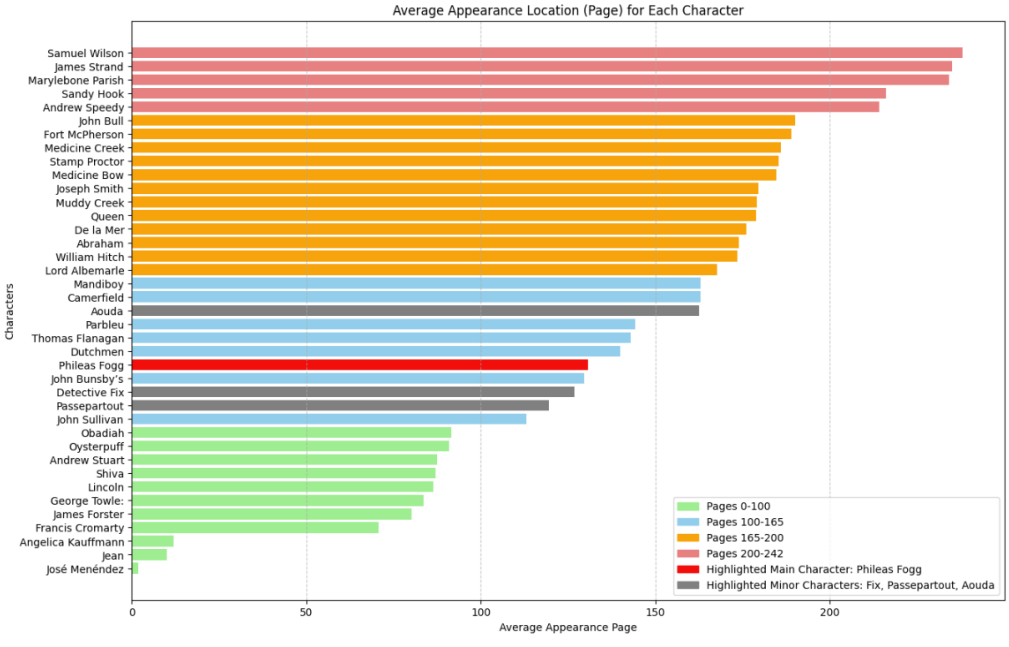

```
Character with the most connections (largest number of arcs): Phileas Fogg with 19 connections
```

3) **Strength of Connections:** I also examined which character had the highest number of strong connections. This was determined by calculating the strength of the relationship between characters based on their co-occurrences within three lines of each other. A character with many strong connections is indicative of their importance and influence within the social network of the story. According to the theory of social network analysis, characters with a higher number of strong connections (high degree centrality) are often central figures within the network, highlighting their key role in the narrative.




```
Character with the highest sum of connections: Phileas Fogg with a sum of 1.03
```

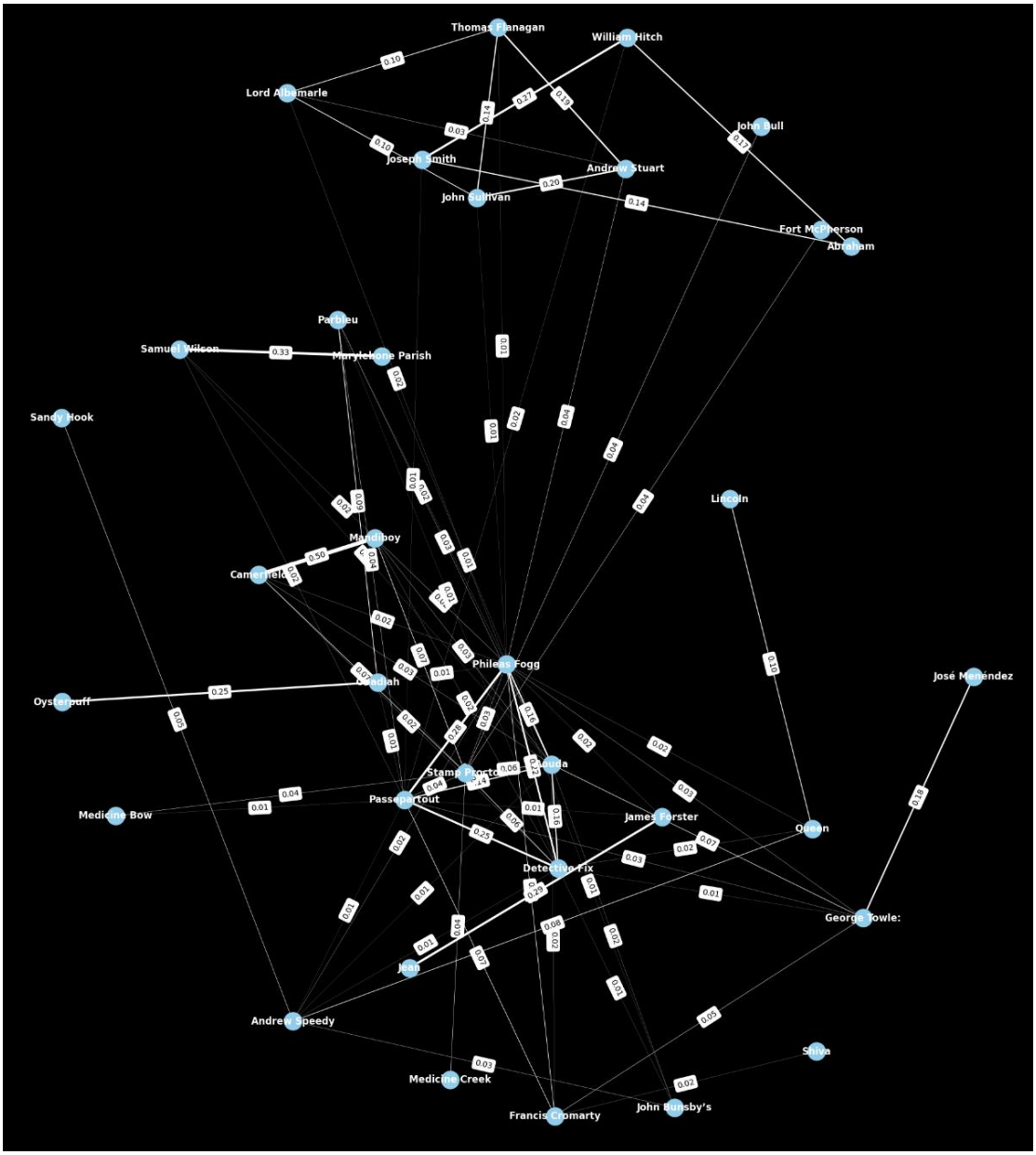
4) **Average Position in the Book:** Lastly, I calculated the average position of each character's appearances across the pages of the book. Characters whose average position is in the central pages of the book are likely to be more central or influential in the narrative. This indicates that these characters are involved in key plot points and interactions throughout the story, appearing consistently from the beginning to the end. Such central placement can mean that the character plays a significant role in the development and progression of the narrative, akin to nodes in a social network that are central and crucial connectors or hubs within the network.



B. <u>Second Question – What are the social circles of the characters?</u>

To answer this question, I tested the **strength of the relationship between the characters in the book**, with the nodes representing the characters and the arcs the strength of the relationship (using function number 6, which counts the number of times the characters appear within three lines of each other and dividing the number of shared appearances by the number of all appearances of the characters together).

In addition, I checked **connected components** and discovered that the book is one connected component that contains all the characters. It can be concluded that all the characters are related to each other in some way in the story.



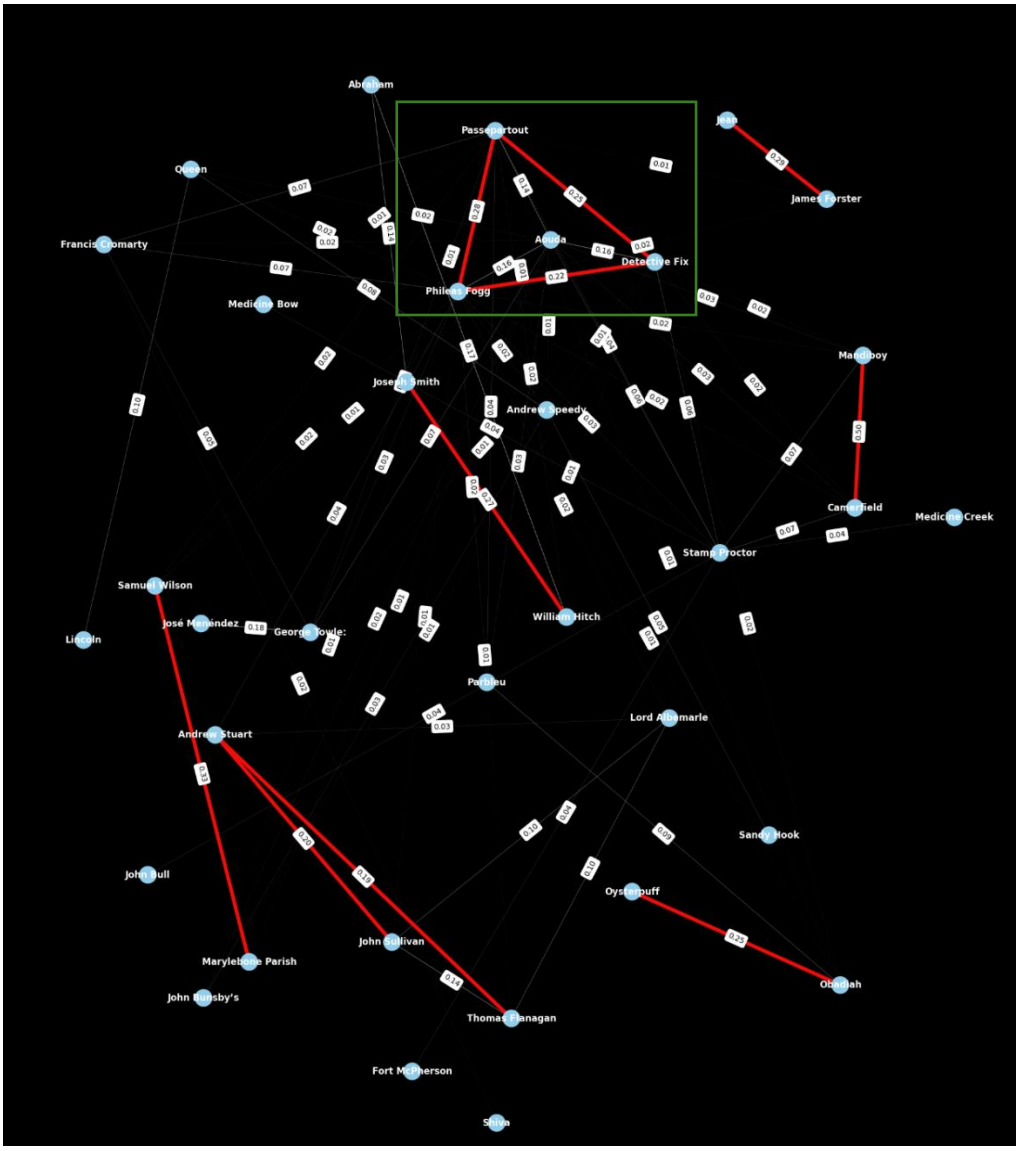

```
Connected Components:
Component 1: {'Shiva', 'Lincoln', 'Camerfield', 'Medicine Creek', 'Jean', 'Parbleu', 'Abraham', 'Oysterpuff', 'José Menéndez', 'Queen', 'Lord Albemarle',
```

Furthermore, I checked whether the network is **directed or undirected** and found that it is undirected, meaning that the relationships between characters are reciprocal. In other words, if character A is connected to character B, then character B is also connected to character A, reflecting the reciprocal nature of their interactions within the story.

```
[78] def check_graph_directionality(graph):
         if isinstance(graph, nx.DiGraph):
             return "Directed"
         else:
             return "Undirected"

     graph_type = check_graph_directionality(R)
     print(f"Is the graph directed? {graph_type}")

     Is the graph directed? Undirected
```

C. **<u>Third Question - What is the strength of the relationship between the characters?</u>**

I addressed this question by identifying and highlighting the 10 strongest connections in the network graph. These connections were sorted based on their weight, which represents the strength of the connection between characters, and were marked in red. <u>It should be noted that the results confirm that the main characters in the book, Fogg, Pix and Passepartout, are significantly related to each other, which emphasizes their central roles and their frequent interactions throughout the story.</u>



Part V- Results

1. **Results -**
   A. **Identifying the main characters** - by counting the number of appearances of each character and analyzing their relationships, I identify the main characters in the book. Phileas Fogg, Detective Fix and Passepartout appeared as the most common characters and were central to the network.
   B. **Social Circle Analysis** - Using connected component analysis, I discovered that the book's network forms a single connected component. This indicates that all the characters are related to each other directly or indirectly. This interconnectedness reflects the narrative in which characters, regardless of their prominence, are part of a wider social network. Additionally, it implies that even minor characters contribute to the storyline and influence the main characters.
   C. **Strength of Relationships** - I calculated the strength of relationships by examining the frequency of co-occurrences within three lines of each other. The strongest connections are identified and highlighted. The strongest connections are between the main characters Fogg, Fix and Paspertau which emphasize the key interactions between them. It reveals the dynamics of alliances and conflicts that are central to the plot.
2. **Limitations –**
   A. Named entity recognition accuracy: The NER process may not perfectly recognize all character names, leading to possible omissions, misidentifications, or duplications between characters.
   B. Co-occurrence window: The chosen threshold of three lines to define connections may not capture all significant interactions, and different thresholds can yield different results.
   C. Minor characters with strong connections: Characters with fewer appearances can still have strong connections with other characters. For example, the characters Camerfield and Mandiboy have a strong connection of 0.5 despite being minor characters. This can occur because their interactions, though fewer, are concentrated and meaningful within the narrative.

Part VI – Business Applications

Business Model for the Software Itself -

The software I developed identifies and analyzes character interactions within a text using various functions and theories from social network analysis. I employed Python libraries to extract and process text from PDFs, detect character names, and determine relationships based on their proximity within the text. This method can be applied across multiple business sectors. For example, human resources departments can use software to analyze internal documents, emails, and reports to identify key personnel and interactions, helping to understand organizational dynamics and improve employee relations. In educational settings, teachers and professors can leverage the tool to analyze texts and identify key themes, characters, and relationships, enhancing teaching materials and classroom discussions. Legal firms and compliance departments can utilize the software to review legal documents, contracts, and case files to pinpoint central figures and their interactions, aiding in thorough legal reviews and compliance checks. Additionally, businesses in the financial sector can use the tool to analyze market reports and articles to identify influential companies and stocks, informing strategic decisions. This software provides valuable insights across various industries, facilitating informed decision-making based on detailed textual analysis.

Targeted Marketing Strategies –

Using insights from the social network analysis of characters allows for more efficient and targeted marketing strategies:

A. **Central Character for Dissemination -** Utilizing the central character, such as Phileas Fogg, for distributing products or messages due to his broad network and significant influence. This approach focuses on targeting a highly influential character to maximize reach and impact.
B. **Avoiding Dissemination in Strong Social Circles** - Avoiding the dissemination of messages through characters within strong social circles of the central character, like Passepartout and Fix, as the message would already reach them through the central character. This ensures that efforts are not wasted on redundant paths.

Targeted Marketing and Product Improvement –

Insights derived from the social network analysis can also be used for enhancing existing products:

A. **Product Development -** Developing new products based on the understanding of relationships and social dynamics. This enables the creation of personalized products that provide better user experiences.

B. **Improvement of Existing Products -** Adapting existing products and services to offer personalized experiences based on social dynamics and interactions. Understanding social connections can help improve existing products to better meet user needs.

Shlomi Shor