Mini Project III – Network Analysis

Ali Kaya
Åbo Akademi University (ÅAU)
ali.kaya@abo.fi

November 2, 2023

Abstract

In this mini project, we harness the power of network analysis, built upon the foundation of Named Entity Recognition (NER), to comprehensively explore the intricacies of the Harry Potter series. Our analysis encompasses diverse dimensions, including co-occurrence patterns of characters, character network construction, community detection, centrality analysis, and tracking the evolution of characters. NER plays a pivotal role in identifying characters and entities within the series, forming the basis for our exploration. By scrutinizing the character relationships, we unveil hidden narrative dynamics, enriching our understanding of the beloved literary masterpiece. This project showcases the potent synergy between storytelling and data analysis, offering a unique perspective on the wizarding world and the endless possibilities that arise when literature meets data science.

1 Introduction

In the ever-evolving landscape of literature and machine learning, an exciting synergy emerges as we embark on a journey to explore the intricate narratives of beloved series like Harry Potter through the lens of network analysis. Literature, a repository of human experiences, emotions, and interconnected narratives, has long captivated readers with its immersive storytelling. Machine learning, on the other hand, equips us with powerful tools to unearth patterns, hidden insights, and quantifiable relationships in vast textual datasets. When these two worlds converge, it unlocks a realm of endless possibilities, paving the way for a deeper understanding of not only the narratives themselves but also the underlying structures and character dynamics.

The Harry Potter series, authored by J.K. Rowling, serves as a prime exemplar of a literary universe that has enthralled millions worldwide. Its rich tapestry of characters, each with a unique role to play, unveils a web of relationships, friendships, rivalries, and allegiances that shape the wizarding world. These intricate interactions have not only captured the imaginations of readers but also provide a fertile ground for exploration using machine learning techniques. Through network analysis, we aim to deconstruct the characters' connections and delve into the narrative dynamics that define this enchanting saga.

In this mini project, we will harness the power of network analysis methodologies and tools to undertake a comprehensive exploration of the Harry Potter series. Our analysis will encompass the following key dimensions:

- Co-occurrence Analysis: Our journey begins with a deep dive into the frequency and patterns of character co-occurrence within the enchanting world of Harry Potter. Through co-occurrence analysis, we aim to unravel the intricate connections between characters, shedding light on the recurring encounters and alliances that drive the narrative.
- Character Networks: We will craft several networks throughout volumes where characters become the nodes, and the ties that bind them manifest as connections. This network-centric approach unveils the underlying structure of character interactions, rendering a visual representation of the web of relationships that shape the wizarding universe.
- Centrality Measures: Through the lens of centrality metrics, including degree centrality, betweenness and closeness centrality, we will pinpoint the characters that stand as pillars of influence and importance within the series. This quantitative exploration allows us to distinguish the protagonists, the unsung heroes, and the pivotal figures who steer the narrative.
- Community Detection: Utilizing sophisticated

network analysis algorithms, we will embark on a quest to unearth the hidden communities or groups of characters who share frequent and meaningful interactions. By identifying these distinct communities, we gain fresh insights into the social fabric of the Harry Potter series, discovering the bonds that transcend the pages.

- Evolution of Characters: The Harry Potter series evolves across its books, and so do the characters and their relationships. Our analysis extends to tracking these transformations, enabling us to witness the ebb and flow of character dynamics across the literary saga. By discerning the evolution of character relationships, we delve deeper into the heart of the narrative.

Through this multidimensional analysis, our objective is to not only uncover the hidden layers of the Harry Potter series but also to showcase the remarkable potential of network analysis in unraveling the complexities of literature. This exploration offers a unique perspective on a beloved literary masterpiece and demonstrates the intersection of storytelling, data, and the endless possibilities that emerge when these worlds converge.

2 Data Preparing and Preprocessing

2.1 Books and Characters

Essentially, the Harry Potter series comprises seven books that we procured online in plain text format. We conducted fundamental content adjustments to meticulously organize and store each of the seven volumes separately, ensuring they are preserved as distinct text files. You can find the books here

At the same time, we obtained a JSON file that encompasses all the characters within the Harry Potter series. This file comprises two key attributes: "Name" and "Description," and you can access the corresponding file here.

2.2 NER for data Preprocessing

Named Entity Recognition (NER) stands as an indispensable element within the data preprocessing workflow. NER methods excel at the detection and retrieval of named entities, encompassing entities like characters' names, organizational references, geographical locations, date expressions, and pertinent data from unstructured textual content. In the context of this project, we leverage NER to extract entities on a per-sentence basis, treating them as discrete tokens. Subsequently, we implement a filtering process using JSON content mentioned in section 2.1 to eliminate superfluous recognized entities, guaranteeing that the entities retained exclusively pertain to characters intricately woven into the fabric of the sentences.

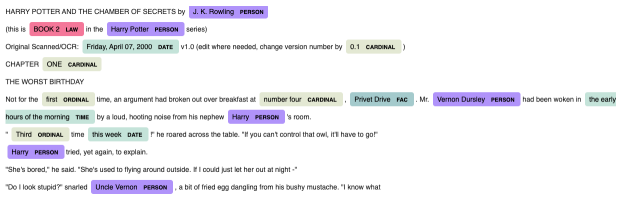

Figure 1: NER sample for Book 2

The whole processes of data preprocessing can be summarized as follow:

- Importing the books in TXT format from a GitHub repository to the Colab runtime storage.
- Parsing each book sentence by sentence and applying Named Entity Recognition (NER) to identify entities. The results are stored in a dataframe, including the sentence, its position in the book, and the recognized entities.
- Employing a JSON file containing genuine character names to filter the entities recognized by NER. The refined data is then organized in a dataframe, capturing the verified characters, their associated sentences, and their respective positions in the book.
- Establishing parameters such as window size and sentence distance to compute co-occurrence between unique characters.
- Aggregating the co-occurrence data for character pairs within each volume and saving the outcomes in dedicated dataframes for further analysis.

In our particular context, the fundamental prerequisite for calculating connections and relationships lies in precisely defining the concept of co-occurrence. It's essential to acknowledge that not every sentence within the text contains characters. For instance, once we've meticulously filtered out the superfluous entities, our data might look something like this: within the initial 20 sentences of a volume, character A appears in the 5th sentence, character B and C appear in the 9th sentence, and character A and B appear together in the 11th sentence, and so forth.

We introduce the concept of "window size" which signifies the number of rows between sentences containing these characters. Additionally, we establish "sentence distance" as the number of rows between the natural order of the sentences. To illustrate with the earlier example, the number of rows between sentences containing characters is only 2, while the number of rows between the natural order of the sentences is 6 (i.e. 11 - 5). These two definitions is used to calculate the co-occurrence between a pair of characters

In this project, we have configured the "window size" to be 3, and the "sentence distance" to be 40. Essentially, this setup implies that for a given pair of characters to establish a co-occurrence, they must be mentioned within a maximum distance of 40 consecutive sentences from each other. Top 5 of the summarized relationship(i.e. characters co-occurrence) per volume can be fund in figure2

```
============Volume 1============
====== Relationship Samples =======
    source   target  value
0   Dudley    Harry    256
1   Dudley      Ted      3
2  Petunia      Ted      3
3   Dudley  Petunia     17
4    Albus    Harry     14
============Volume 2============
====== Relationship Samples =======
    source   target  value
0    Harry   Vernon     11
1  Petunia   Vernon      5
2   Dudley    Harry     82
3   Dudley   Vernon      9
4   Dudley  Petunia     15
============Volume 3============
====== Relationship Samples =======
     source    target  value
0  Bathilda     Harry      2
1    Dudley     Harry     60
2     Harry       Ron   1131
3       Ron   Vernon      8
4     Harry  Hermione    464
============Volume 4============
====== Relationship Samples =======
  source  target  value
0  Harry    Rita    171
1  Frank   Harry     26
2  Frank    Rita      1
3  Frank  Nagini     10
4  Harry  Nagini     16
============Volume 5============
====== Relationship Samples =======
    source   target  value
0    Harry   Vernon     38
1   Dudley    Harry    235
2   Dudley   Vernon     21
3    Harry  Petunia     16
4  Petunia   Vernon      9
============Volume 6============
====== Relationship Samples =======
      source     target  value
0     Amelia  Emmeline      1
1   Narcissa    Severus      8
2  Bellatrix    Severus     17
3  Bellatrix  Narcissa     26
4      Harry  Narcissa     20
============Volume 7============
====== Relationship Samples =======
      source   target  value
0      Harry     Pius      6
1      Harry    Lucius     69
2     Draco    Lucius     20
3  Bellatrix    Draco     22
4  Bellatrix  Narcissa      9
```

Figure 2: Top 5 relationship per volume

3 Network Analysis

3.1 Visualization of Networks

In an undirected graph where characters represent nodes and relationships form edges, we have the ability to create networks based on the provided data. These networks can be visually represented, with node size determined by their degree. The degree of a node corresponds to the number of edges connected to it, serving as a metric to gauge its connectivity within the graph.

We employ NetworkX to generate the graph object, but for visualization, we opt for Pyvis.network due to its exceptional visual impact.

Pyvis stands out for its ability to generate visually stunning, interactive network visualizations. It offers diverse layout options, customization features, and real-time interaction capabilities that enable users to delve into the nuances of network structures with ease. Here are the networks we generated for volumes one to seven.

From the generated networks in figure 3 4 5 6 7 8 9, we can easily tell the following observations:

- The character network in every single volume is complicated
- Harry Potter is always the main character in this series
- Some of the characters only exist in some volumes
- The evolution of the networks between networks(i.e. volumes) is significant

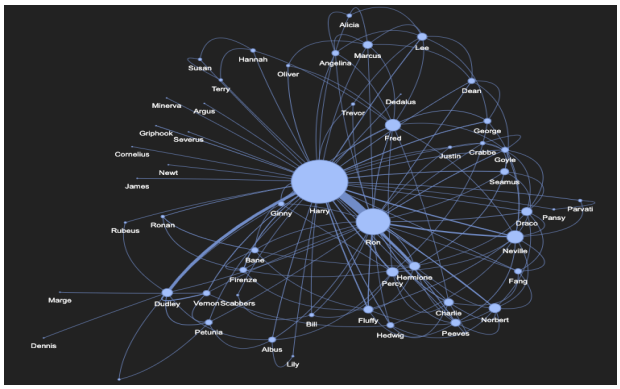

Figure 3: Networks for book1

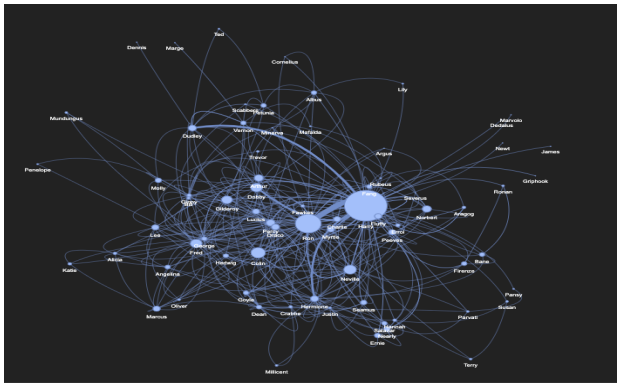

Figure 4: Networks for book2

Figure 5: Networks for book3

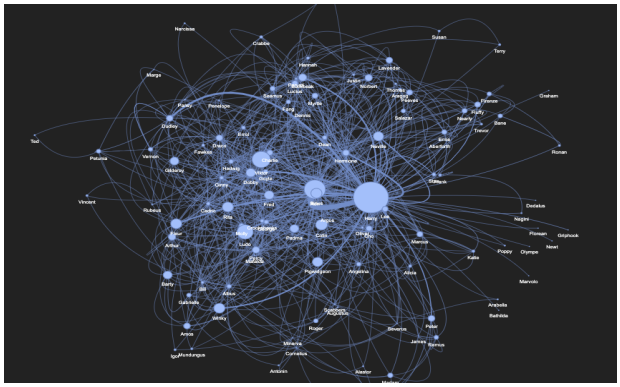

Figure 6: Networks for book4

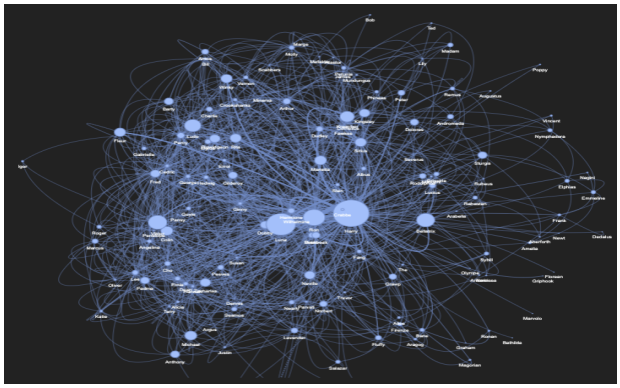

Figure 7: Networks for book5

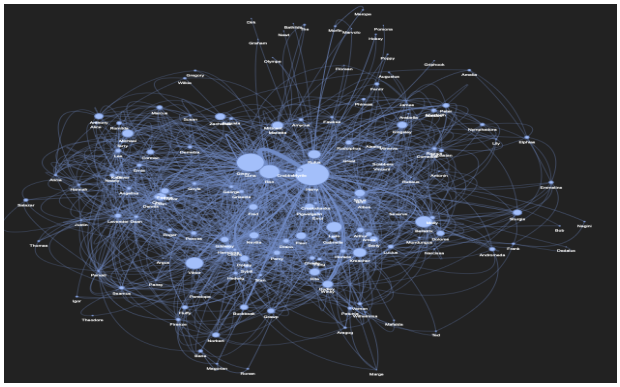

Figure 8: Networks for book6

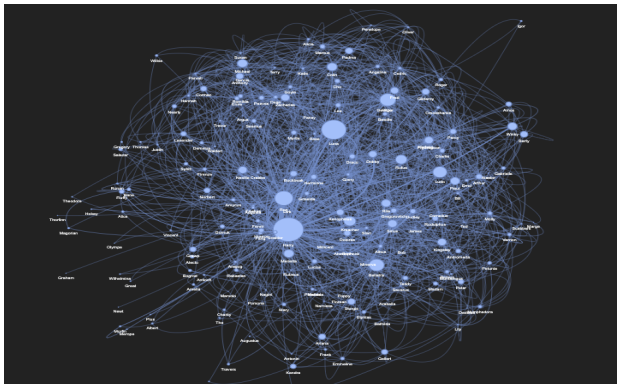

Figure 9: Networks for book7

3.2 Centrality Measures

Degree centrality is a fundamental concept in network analysis, serving as a key metric to gauge the importance of nodes within a network. It measures the number of direct connections a node has to other nodes, making it a fundamental indicator of a node's influence and prominence in the network. In essence, degree centrality is a reflection of a node's popularity or how many "neighbors" it has. Nodes with high degree centrality often play pivotal roles in information dissemination, communication, or the flow of resources within the network. As a result, degree centrality provides essential insights into the structural dynamics of a network, helping us identify key players and their impact on network behavior.

Betweenness centrality is a critical network metric that quantifies the significance of a node in terms of controlling the flow of information or resources between other nodes in the network. This metric is particularly valuable in identifying nodes that act as "bridges" or "gatekeepers" within the network, as they play a pivotal role in facilitating communication and interactions among different parts of the network. Nodes with high betweenness centrality have the potential to influence the efficiency of information transfer and overall network connectivity. They act as critical intermediaries in ensuring the smooth and effective functioning of the network, making betweenness centrality a powerful tool for understanding the network's structural and functional dynamics.

Closeness centrality is another significant measure in network analysis, focusing on the accessibility and efficiency of information or resource transfer for a specific node. This metric assesses how quickly a node can reach all other nodes in the network, highlighting nodes that can efficiently interact with their neighbors and beyond. Nodes with high closeness centrality are located in positions that allow for rapid communication and resource exchange, making them essential for the network's cohesion and responsiveness. Closeness centrality offers insights into the effectiveness of nodes in terms of information dissemination and their capacity to influence the network's overall speed and efficiency. It serves as a crucial tool for identifying nodes that can efficiently relay information or resources to various parts of the network, contributing to its overall robustness and functionality.

We graphically(Figure 10 11 12 13 14 15 16) represent the nine most significant characters by employing Degree centrality, Betweenness centrality, and Closeness centrality in parallel for each volume. From this analysis, we can make the following observations:

- While each of the three centrality methods has its unique focus, it's noteworthy that Harry

Potter consistently ranks as the most significant character, and this aligns with the overarching theme of the series, which revolves around his adventures.

- A substantial contrast in the importance of Harry Potter relative to other characters arises when applying the Betweenness centrality method, as compared to other centrality measures. This contrast underscores the fact that Harry Potter serves as the primary and exclusive "bridge" within each network (i.e., volume), playing an indispensable role in facilitating the entire narrative.

- When comparing Closeness centrality to Degree centrality, it becomes evident that the centralities of characters other than Harry Potter gain prominence. This observation aligns with the rationale that Closeness centrality amplifies the significance of supporting roles, especially if they engage in substantial interactions with the protagonist. This accentuates the fundamental principle of Closeness centrality, where high Closeness centrality positions are strategically located for swift communication and efficient resource exchange, making them pivotal for network cohesion and responsiveness.

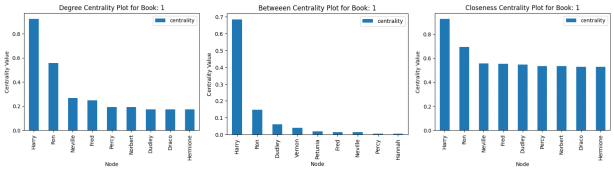

Figure 10: Centrality Book1

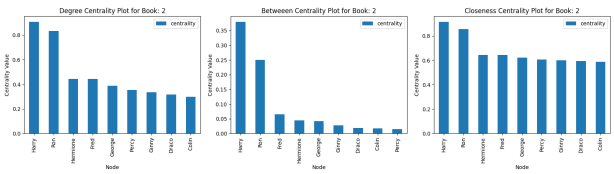

Figure 11: Centrality Book2

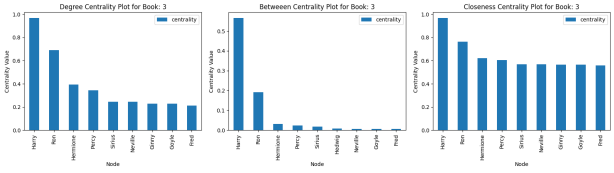

Figure 12: Centrality Book3

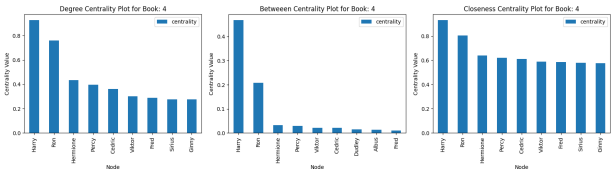

Figure 13: Centrality Book4

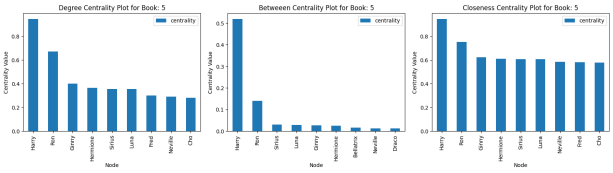

Figure 14: Centrality Book5

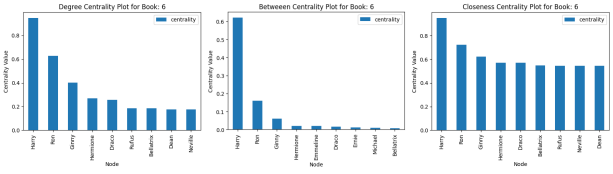

Figure 15: Centrality Book6

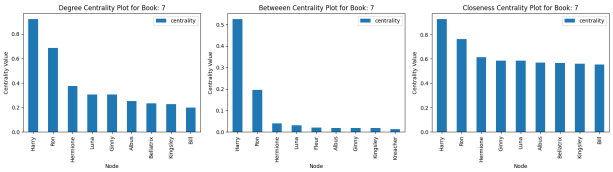

Figure 16: Centrality Book7

3.3 Community Detection

Community detection is a fundamental task in network analysis, and the Louvain method is a popular and efficient approach for finding communities in complex networks. It was introduced by Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre in their paper "Fast unfolding of communities in large networks" in 2008. The Louvain method is a heuristic algorithm that aims to identify communities or clusters within a network by optimizing a quality function.

The Louvain method is a bottom-up and greedy approach to community detection. It works in two phases:

Phase 1 (Modularity Optimization): In this phase, the algorithm tries to maximize the modularity of the network, which is a measure of the quality of the community structure. Modularity quantifies the difference between the actual number of edges within communities and the expected number of edges in a random network with the same node degrees. The Louvain method iteratively merges or "unfolds" communities to improve modularity.

Phase 2 (Refinement): After finding an initial community structure, the Louvain method iteratively refines the community assignment. It considers each node as a separate community and tries to improve modularity by moving it to a neighboring community, provided that the move increases modularity. This process continues until no further improvement can be made.

Figure 17 in plain text and Figure 18 19 20 21 22 23 24 in graphs are the communities detected for complex networks in Harry Potter series using Louvain method:

Figure 17: Community Samples in Plain Text

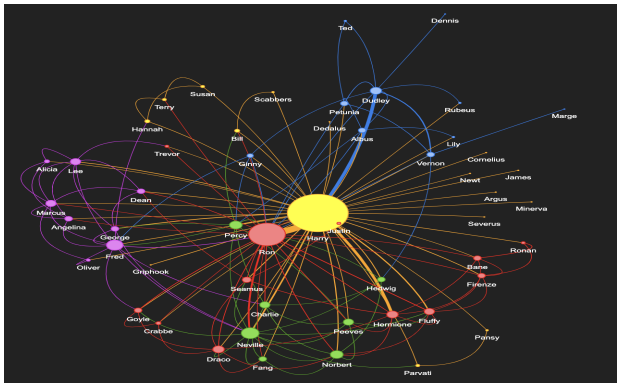

Figure 18: Communities for book1

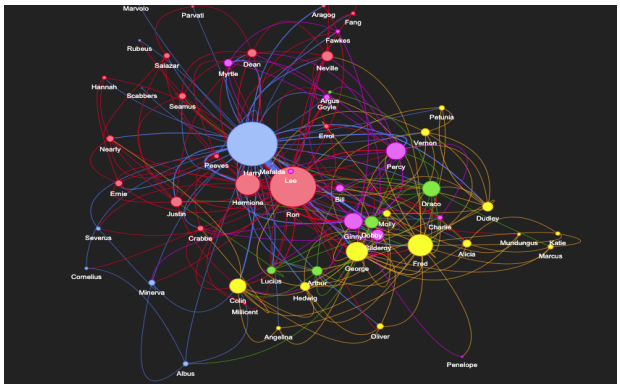

Figure 19: Communities for book2

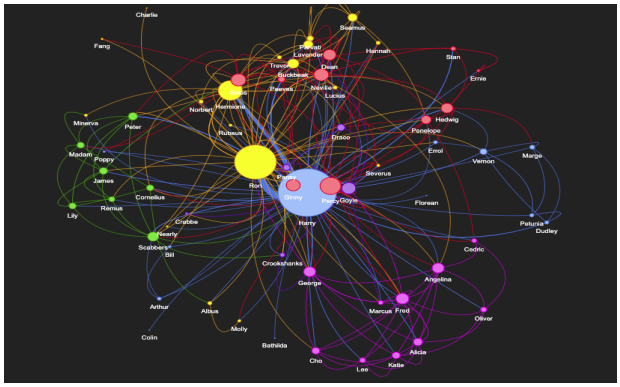

Figure 20: Communities for book3

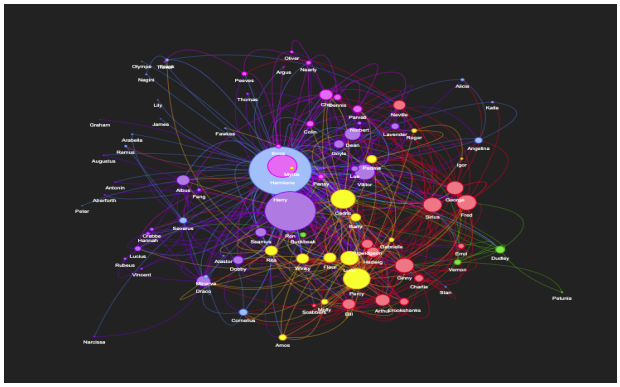

Figure 21: Communities for book4

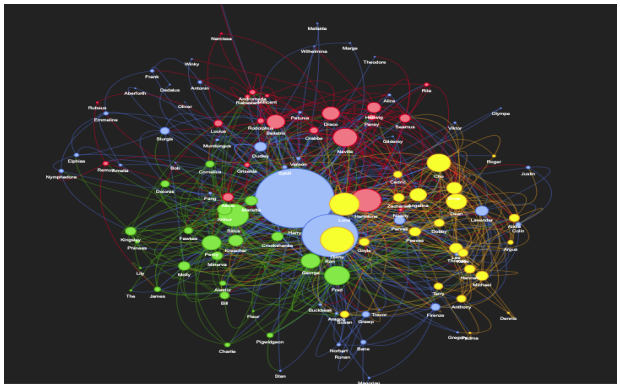

Figure 22: Communities for book5

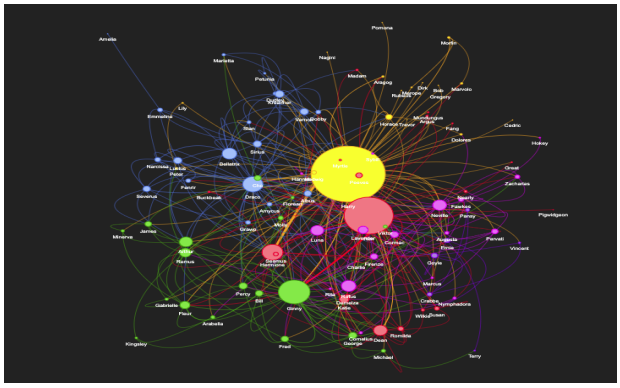

Figure 23: Communities for book6

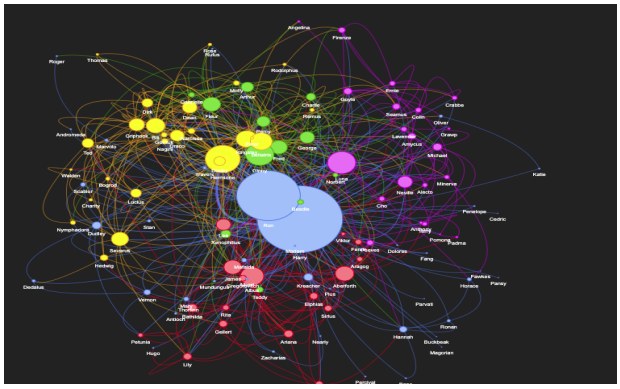

Figure 24: Communities for book7

3.4 Evolution of Characters

Character evolution in a series of novels is a complex and multifaceted phenomenon. It is often essential for advancing the plot and developing the narrative. As the story progresses, characters face new chal-

lenges, make decisions, and undergo personal growth, all of which contribute to the narrative's complexity and depth.

In this project, we employ degree centrality as the foundational metric to gauge the progression of characters within the series. We implement two distinct evolution timelines:

- The top 6 Characters Evolution Over Time based on the summation of the degree centrality among all the series. This will highlight the most influential characters over time.
- The top 6 Characters Evolution Over Time based on the deduction of maximum degree centrality and the minimum degree centrality among all the series. This will highlight the top 6 fastest changing characters over time.

We could tell from figure 25 that Harry Potter holds the highest degree of influence in the entire series, with Ron Weasley, the youngest son of Arthur and Molly Weasley who is best friends with Harry Potter and Hermione Granger, following closely as the next influential character.

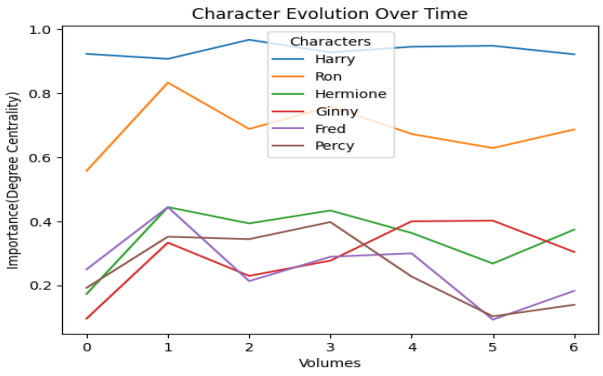

Figure 25: The top 6 influential characters over time

We could also tell from figure 26 that the depiction of George Weasley began in volume 2, only to fade away after the events of volume 5. In fact, there are lots of discussion where George Weasley is after book 5 - "Order of the Phoenix" even in IMDB.

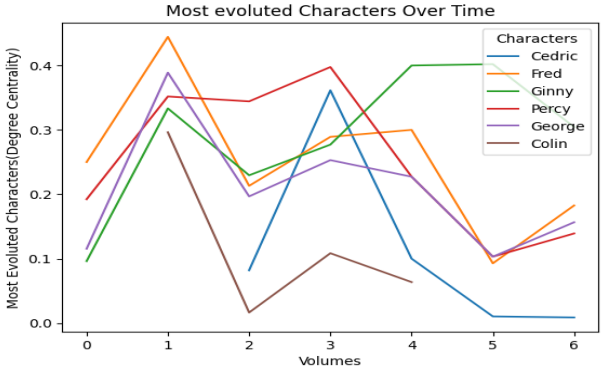

Figure 26: The top 6 fastest changing characters over time

4 Conclusion

In the enchanting world of the Harry Potter series, we embarked on a journey where the magic of storytelling met the precision of data science. Through the lens of network analysis, we delved into Co-occurrence Analysis, Character Networks Visualization, Community Detection, Centrality Measurements as well as Character Evolution. A significant challenge in this project lies in utilizing Named Entity Recognition (NER) to tokenize the entities, which can be time-consuming and may yield relatively lower accuracy. Fortunately, the pre-extracted characters provide a valuable reference that enables us to cross-reference and validate the entities extracted from the volumes using NER. It's worth noting that this approach may not be applicable to other projects without external reference data, and in such cases, more sophisticated methods would be required as needed.

As we conclude this project, it is evident that network analysis is not just a tool for unraveling complex structures; it is a key to unlocking the hidden depths of literature. The characters of Harry Potter, including Harry himself, Ron, Hermione, and countless others, are more than fictional beings; they are vessels of profound storytelling. Network analysis has allowed us to appreciate their evolution, their interconnections, and the impact of their journeys in a new light.

This project stands as a testament to the infinite possibilities that arise when literature and data science converge. The magic of storytelling and the precision of data analysis are not exclusive realms; they are intertwined, offering new horizons for exploration. As we close the book on this analysis, we find that the characters of Harry Potter, their relationships, and the stories they inhabit are not just words on a page; they are a testament to the timeless power of storytelling, waiting to be discovered through the lens of data and network analysis.