Algorithm for Massive Data - Link Analysis

Tommaso Premoli - 34221A

I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work, and including any code produced using generative AI systems. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

Abstract

The aim of this project is to analyze the network of connections between books reviewed by Amazon users through link analysis and network analysis techniques. The purpose is to identify the most central and influential books within the network in order to understand which works act as hubs, sharing a large number of readers with other books, and which occupy peripheral positions, attributable to thematic niches. To achieve this goal, various measures of centrality are used, including PageRank, which are useful for quantifying the importance and influence of individual books in the network. One of the practical objectives of this analysis is to improve recommendation systems by suggesting books that occupy central positions in the graph and are therefore closer to a wide range of user interests.

Contents

1 Introduction

User book reviews allow you to generate a network of connections that links the volumes reviewed by the readers themselves. Analysing this network allows you to understand the structure of the relationships between books, identifying those that occupy central positions and are therefore more influential or shared within the reader community. Starting from this network, it is possible to develop recommendation systems that can suggest new books not only based on general popularity, but also on the collective preferences of users. In simple terms, if a reader has reviewed a particular book, they are likely to appreciate another book reviewed by the same users.

This recommendation or popularity rating mechanism is achieved by computing various measures of centrality, with a particular focus on PageRank. PageRank was originally developed by Google to rank web pages, as this measure evaluates the importance of a node not only based on the number of connections, but also on the importance of the nodes to which it is connected. The aim was therefore to rank pages based on their structural relevance in the network of links, rather than on simple textual content. PageRank is based on the idea of the "random surfer", which is a user who browses randomly from one page to another by following the available links: the probability of visiting a given page therefore depends on the number and weight of incoming connections. A key element of the algorithm is the taxation factor, which assigns a small probability at each step of randomly jumping from one node to another in the network, thus avoiding problems related to isolated nodes.

In the context of this project, PageRank allows us to identify the most central and influential books, as they share readers with other books of equal relevance. In other words, a book achieves a high PageRank not only because it is widely

reviewed, but because it is linked to works that are themselves central to the network, thus assuming a key role in the system of reader preferences.

2 Dataset Description

The 'Amazon Books Review' dataset, available on Kaggle and downloaded via public APIs, was used to perform this analysis. As described above, the dataset contains book reviews written by Amazon platform users, including the book title, user ID and rating for each review. Apache Spark was used for data modelling and management, which allows large volumes of information to be processed efficiently. After downloading the compressed file *(.zip)*, two main datasets were loaded:

- **reviews**: which contains the reviews and ratings assigned by users to different books;
- **book_info**: a table containing the main metadata relating to the books, such as author, genre, publication date and link to the product page.

The reviews dataset forms the main basis of the study, as it represents the behavioural component of users. It comprises 3 million rows. Through a join on the book title, information from the book_info dataset was also integrated, in particular the literary genre, considered a useful variable for any specific and in-depth analyses.

2.1 Data Cleaning

Various data cleaning and preparation strategies were adopted in order to ensure the quality and consistency of the dataset. In particular, measures were taken to reduce the number of rows, normalise book titles and finally select a representative sample of the final dataset to be used for subsequent analysis phases.

2.1.1 Row Reduction Strategies

First of all, in order to reduce the number of rows in the dataset and obtain more consistent and meaningful results, we chose to apply targeted filtering strategies rather than simply performing random sampling from the outset.

Two main reduction strategies were adopted:

- **Filtering the most active users** : Only reviews from users who had written at least six reviews were retained. This choice was motivated by the desire to include only users with a higher level of experience and involvement in book evaluation. In this way, the reviews considered are on average more reliable and representative, contributing to a better overall data quality.

- **Filtering the most popular books** Only books that received more than 80 reviews in total were retained. This filter was introduced to exclude lesser-known or poorly reviewed titles, focusing the analysis on books that are popular and relevant to the public.

Applying these criteria reduced the dataset to a total of 526,802 reviews. This reduction allowed for a more focused and meaningful analysis, while improving the quality of the results and computational efficiency in the subsequent stages of processing.

2.1.2 Title Normalization

The main objective of this phase of the analysis is to ensure the consistency and uniqueness of book titles, so that centrality measures (such as PageRank) and other network metrics can be calculated correctly on the title itself. It was decided to base the measures on the book title rather than the identification code in order to obtain results that are more interpretable and meaningful from a visual and analytical point of view. During the preliminary phase, it emerged that numerous books, while referring to the same volume, had variations in their titles due to differences between editions, punctuation, formats (for example, audiobooks), or the presence of additional non-meaningful words such as "illustrated", "volume", "classic edition", etc. This heterogeneity produced duplications and distortions in the analyses, as the same book could appear in different nodes of the graph, generating different PageRank values and compromising the validity of the analysis.

To resolve this issue, four consecutive stages of standardisation were adopted.

1. **Preliminary cleaning of the title text** : In the first phase, the titles were converted to lowercase and cleaned up to remove multiple spaces, punctuation and special characters. The aim of this step was to standardise the titles and reduce formal differences between similar texts, facilitating comparison in subsequent phases.

2. **Removal of non-significant words** : A semantic filter was then applied, eliminating recurring but non-informative terms that caused non-substantial variations between titles referring to the same book. These included words such as edition, illustrated, volume, classics, school library, audiobook, etc. The removal of these terms made it possible to achieve greater consistency between titles, reducing lexical discrepancies due to editorial elements.

3. **Unification through textual similarity (Fuzzy Matching)** In this phase, a lexical similarity approach based on the fuzzy matching algorithm was used, which measures how similar two text strings are to each other. Specifically, the *fuzz.token_set_ratio()* function from the *FuzzyWuzzy* library was used, which compares two strings by comparing sets of words regardless of order and calculating a similarity percentage (0–100). For

each title in the dataset, other titles with a similarity of 90% or higher were searched for. Similar titles were grouped together and represented by a canonical title, chosen from among the longest ones. Although this method significantly reduced duplication, it was not entirely effective in handling more profound variations in meaning or linguistic errors, which is why a more advanced algorithm was chosen.

4. **Semantic clustering with embedding models** : For more accurate normalisation, an approach based on semantic representation models (embedding) was adopted. Each title was converted into a numerical vector representing its semantic meaning using the SentenceTransformer "all-MiniLM-L6-v2" model, which is capable of capturing linguistic and conceptual relationships between texts. Subsequently, the vectors obtained were grouped using the unsupervised clustering algorithm *HDBSCAN*, which automatically identifies groups of semantically related titles. Each cluster represents a set of titles that describe the same book or very similar concepts. Within each cluster, the shortest and most representative title was selected as the canonical title. Titles not assigned to any cluster were considered outliers and kept unchanged. The final result is a complete mapping between original titles and canonical titles, integrated into the dataset via a join in Spark.

2.1.3 Random Sample

As a final step before conducting the actual analysis, a sample comprising 35% of the total data set was extracted. This resulted in a final working data set of 85,067 observations, ensuring a balance between computational efficiency and representativeness.

An example of the final dataset that will be used for analysis is as follows.

	book_title_ds	user_id	profile_name_ds	score_nm	category_ds
0	casino royale bullseye	ALH1W9IJQQH2E	James H. Felder	4.0	['English fiction']
1	the hound of the baskervilles signet	A3MV1KKHX51FYT	Acute Observer	5.0	['Fiction']
2	plainsong	A2NS4ZOVB28DK4	Maria Atas	3.0	['Fiction']
3	the butlerian jihad legends of dune book 1	AFYU5LXHZO5M3	Vilbs "vilbs"	5.0	['Art']
4	great expectations signet	A1DKQCC90FY7NX	mzgambler "Mz"	5.0	['Fiction']
5	charles dickens pickwick papers	AEMZRE6QYVQBS	David A. Baer	5.0	['England']
6	charles dickens pickwick papers	A1Z07UXQ1ZH07V	Judith Kilpatrick	2.0	['England']
7	rainbow six	A2FS82KFC4NUNR	Cai Yixin Jeremy	5.0	['Games & Activities']
8	red rabbit	A2FT22FKKY7VRK	Paul M. Gunther	1.0	['Intelligence officers']
9	lord of the flies a novel	A1LUM076RXMFNV	Beverly J. Scott	5.0	['Fiction']
10	lord of the flies a novel	ABAHAEG2TYIPW	Kevin Bell	5.0	['Fiction']
11	the fountainhead	A3OBW7ZP7B1Y1N	T. VanPool	5.0	['Architects']
12	little women junior	A1MI38S0X74WZT	Jirehh	5.0	['Juvenile Fiction']
13	economics in one lesson pocket	A1IOR3OTTUK3O5	Jeffrey Roberts	5.0	['Biography & Autobiography']
14	satanic verses	A21XQ98KGFJBX8	Sarah "aggie06"	5.0	['Religion']

Figure 1: Example of the final dataset

3 Exploratory Data Analysis

This part of the analysis aims to examine some distinctive features of the dataset in order to better understand its structure and dynamics. As a first step, a bar chart was created showing the distribution of ratings assigned by users to books. From observation of the graph, it emerges that most ratings tend towards very high values, indicating a general propensity among users to express positive judgements about their reading.

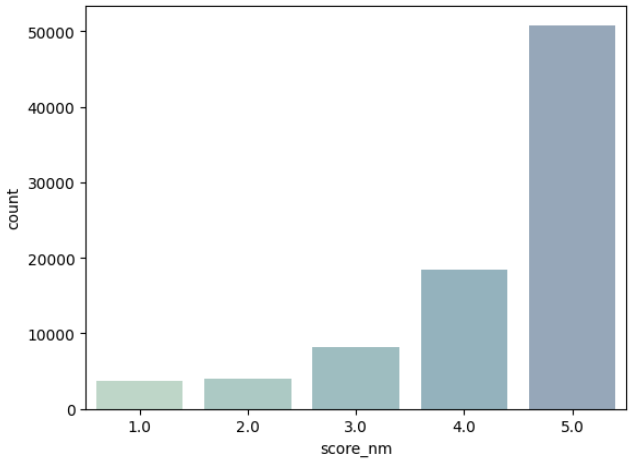

Figure 2: Rating Distribution

Another analysis that was conducted was to look at the books with the most user reviews. The analysis shows that 'The Hobbit or There and Back Again' is the most reviewed book, with over 4,900 reviews, followed by "Ringworld" and 'Jane Eyre'. The ranking shows a prevalence of literary classics, indicating a strong concentration of user interest in iconic and universally recognised titles.

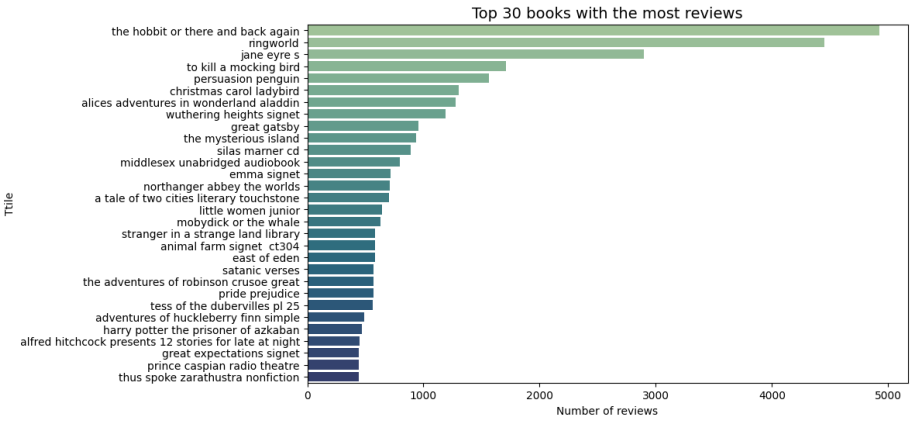

Figure 3: Top 30 books with the most reviews

4 Final Analysis

4.1 Network Construction

As illustrated in the previous chapters, the aim of this study is to evaluate the importance and influence of books on the Amazon platform through the application of centrality measures, such as PageRank. The first step, which we have already completed, consists of constructing the basic dataset necessary for creating the network in which each row represents the association between a user and the book they reviewed. Next, a new dataset was generated consisting of three columns: the first book, the second book, and a weight column, which indicates the number of connections between the two, that is, how many times a user has reviewed both books. This last column therefore represents the weight of the arc in the network.

The exampla is the following.

book_i	book_j	weight
a tale of two cities literary touchstone	great gatsby	282
jane eyre s	wuthering heights signet	273
persuasion penguin	wuthering heights signet	240
jane eyre s	northanger abbey the worlds	235
ringworld	the hobbit or there and back again	230

The resulting graph is structured as follows:

- Nodes: each node represents a book in the dataset;
- Arcs: an arc connects two books if they have been reviewed by the same user;
- Weights: as seen before, the weight of each arc indicates how many users have reviewed both books.

Each node has also been associated with the characteristics of the book, such as its literary genre.

4.2 Network Analysis

The final network consists of 667 nodes and 71,142 arcs. The main measures of this book network are as follows:

Table 1: Network Summary Statistics

Metric	**Value**
Number of Nodes	667
Number of Edges	71,142
Average Degree	213.32
Density	0.32

The average degree of the network is approximately 213.32, which means that each book is connected to an average of over 200 other books, that is on average, more than 200 users have reviewed pairs of books connected within the network. Furthermore, the density is 0.32. This means that the books are fairly interconnected and that 32% of the possible connections between the different nodes are present.

After performing all the steps correctly, we are ready to calculate the centrality measures with a greater focus on PageRank. The calculation gives the following result:

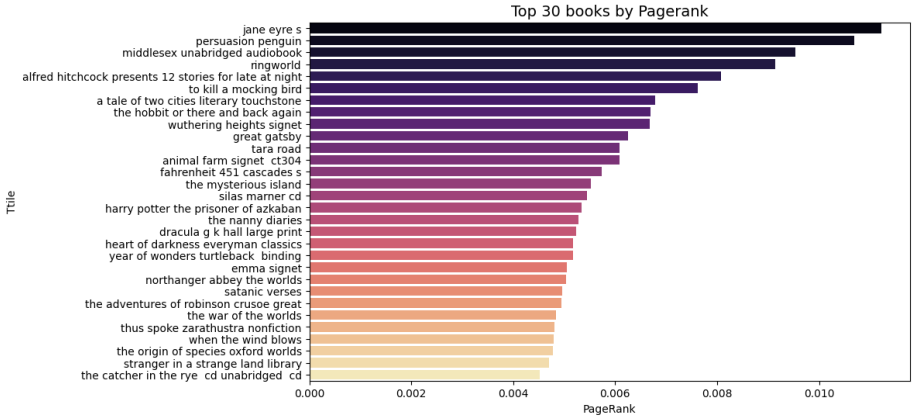

Figure 4: PageRank Result

PageRank analysis shows that the most important books in the network are mainly literary classics, such as Jane Eyre, Persuasion, A Tale of Two Cities and The Hobbit. The presence of titles such as Ringworld and Middlesex also highlights an interest in science fiction and modern fiction. PageRank reflects

not only a book's popularity in terms of reviews, but above all its connection with other central books in the network. In this sense, the titles with the highest values are those most shared among different reader communities, acting as real hubs between different genres and preferences.

In addition, further measures of centrality were calculated to obtain an overall view of the importance and influence of books within the network. Among these, degree centrality is the most immediate measure, as it indicates the number of direct connections each node has with the others. A high degree centrality value suggests that the book is widely co-reviewed with many other titles, thus acting as a hub between different works and reader communities. The result obtained is as follows:

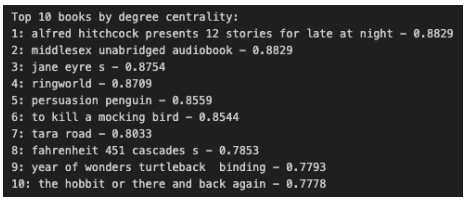

Figure 5: Top 10 books by degree centrality

The output confirms that certain books, such as Middlesex, Jane Eyre and The Hobbit, represent true hubs within the review network, ranking among the most central and connected titles. Finally, a graphical representation of PageRank was created, showing the books with the highest values for this metric. For better readability and interpretation of the graph, only the 100 books with the highest PageRank were considered. In the graph, the books positioned at the centre of the network correspond to those that act as main nodes or hubs, that is, the most influential and connected titles. In contrast, books located in peripheral areas have fewer connections with others and are therefore less central to the overall structure of the network.

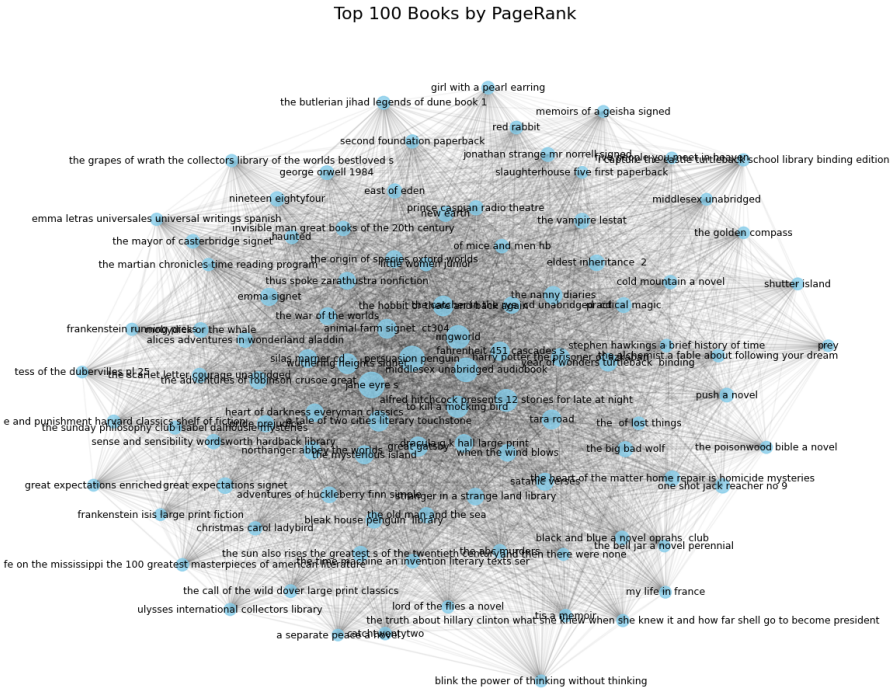

Figure 6: Network visualization

5 Conclusion

This project aimed to analyse the network of connections between books reviewed by Amazon users using link analysis techniques. By calculating centrality measures, with a particular focus on PageRank, it was possible to identify the most influential books in the network, such as Jane Eyre, Persuasion and The Hobbit. The results allow us to identify influential books and shared interests, which provides a better understanding of reader behaviour on the platform.