UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

BEATRIZ DE CAMARGO CASTEX FERREIRA

Modeling thematic relationships in literature through complex networks and similarity analysis

São Carlos
2025

RESUMO

Este estudo desenvolve um modelo computacional para a análise de relações literárias utilizando redes complexas e aprendizado de máquina. O trabalho propõe que a similaridade literária, tradicionalmente avaliada por especialistas, pode ser representada quantitativamente como uma rede em que cada nó corresponde a uma obra e cada aresta expressa proximidade semântica derivada de metadados textuais. Dados bibliográficos do Goodreads, enriquecidos com informações do Wikidata, foram processados por meio de uma abordagem híbrida de extração de palavras-chave (TF–IDF + RAKE) e codificados com Sentence-BERT embeddings. As similaridades par-a-par foram calculadas via similaridade de cosseno, resultando em um grafo ponderado e não direcionado construído no Neo4j. Aplicou-se o algoritmo Leiden para detecção de comunidades, permitindo identificar agrupamentos correspondentes a gêneros e subgêneros literários. Dois conjuntos foram testados: uma rede multilíngue com 5000 obras e uma rede em inglês com 10000 obras. Varreduras paramétricas determinaram valores ótimos (limiar = 0,55; top-k = 25; $\alpha = 0,11$) que equilibram cobertura e modularidade. Os resultados demonstram que representações baseadas em redes complexas capturam organização temática emergente: gêneros e estilos literários surgem como comunidades semânticas densas. A metodologia estabelece uma ponte entre modelagem quantitativa e análise literária, oferecendo uma base reprodutível para futuros sistemas de recomendação e visualização.

Palavras-chave: redes literárias; similaridade semântica; detecção de comunidades; grafos de conhecimento; humanidades digitais.

1. INTRODUCTION

1.1 CONTEXT AND PROBLEM STATEMENT

Over the centuries, humanity has accumulated vast amounts of cultural information, much of it in analogue form. Although digitization has improved access, the sheer scale makes it impossible for any individual to engage with more than a fraction of it. The landscape for books and other

print-based works remains fragmented, inconsistent, and poorly optimized for online discovery [9, 22].

Traditional cataloging systems—such as library classifications and subject taxonomies—still reflect nineteenth- and twentieth-century conceptions of knowledge. Rebuilding them would require recataloging entire collections, a task both logistically and financially prohibitive. As a result, users outside those systems struggle to find new or unexpected connections between works.

This limitation parallels earlier challenges in film and music, where large portions of content remained obscure until machine-learning-based recommender systems transformed access. Similar hybrid approaches can, therefore, be adapted to cultural and literary analysis [1, 20].

1.2 THEORETICAL BACKGROUND AND RELATED WORK

Network theory offers a mathematical framework for representing complex systems as relational structures, where entities are nodes and their interactions are edges. Such models reveal global patterns like modularity and small-world organization [27, 3, 14], uncovering relationships invisible in hierarchical taxonomies.

Modeling long-form texts poses a central challenge: defining edges that capture the many dimensions of semantic similarity. Advances in representation learning now allow textual or categorical data to be embedded in high-dimensional vector spaces, enabling numerical comparison of meaning [4, 20]. When combined with graph structures, these embeddings form hybrid knowledge graphs that integrate relational topology with semantic context.

These methods—blending machine learning and network analysis—have been effectively applied to recommend, cluster, and visualize cultural data [25, 22, 9, 8], revealing thematic communities across large literary corpora [5].

1.3 OBJECTIVES AND SCOPE OF WORK

This study constructs a knowledge network of literary works, combining methods from complex network analysis and machine learning to uncover latent connections and patterns of similarity between texts. The pipeline begins with the Goodreads dataset [6, 7], linking works,

editions, authors, series, and related metadata. These records are normalized and enriched with Wikidata [28], and keywords are extracted from textual descriptions.

Subsequently, embedding and similarity computations model semantic relationships between works, producing a weighted similarity network. Finally, community detection identifies clusters whose qualitative characteristics correspond to literary genres and thematic groupings.

1.4 DATASET AND TOOLS

This work uses the *Goodreads dataset* curated by the **UCSD Recommender Systems Group** [7, 25, 26], enriched with additional metadata retrieved from *Wikidata* [28].

The graph structure is stored in and queried through the **Neo4j** graph database [12], while analytical operations use the **NetworkX** [13] and **igraph** [8] libraries. Machine-learning and numerical operations employ **scikit-learn** [20], **NumPy** [15], and **pandas** [16]; embeddings are computed with **Sentence-Transformers** [21].

The remainder of this text is organized as follows: Section 2 presents the materials and methods, Section 3 reports the results and discusses their implications, and Section 4 summarizes conclusions and outlines possible extensions.

All source code, data, and documentation for this project are available in the following GitHub repository: <https://github.com/BeatrizCastex/literary-knowledge-network>. Information on how to utilize the code archive, mathematical definitions of the algorithms and measures used and samples of results can be found in the appendices: <https://tinyurl.com/4t373djn>.

For transparency, the author discloses the use of OpenAI's GPT-5 model for literature exploration, code assistance and debugging, qualitative cluster categorization, table rendering, text review and code documentation. All algorithmic decisions, data processing, and analyses were designed, executed, and validated by the author.

2 METHODS

2.1 OVERVIEW OF THE METHODOLOGICAL FRAMEWORK

This study combines methods from complex networks, information retrieval, and natural language processing to construct and analyze a literary knowledge network. Figure 1 summarizes the modular workflow, which proceeds through seven stages: data extraction and normalization,

enrichment, keyword processing, embedding generation, similarity computation, community detection, and evaluation/visualization.

Each stage transforms the representation of a literary work—from descriptive text to structured metadata, from metadata to numerical vectors, and from vectors to a weighted relational graph. The data preparation phase involves selection, normalization, and enrichment through keyword extraction and external knowledge integration; the graph construction phase organizes entities and relations in Neo4j; and the analysis phase computes similarities and applies community detection to identify clusters reflecting literary genres and thematic affinities.

This approach follows principles of complex-systems modeling [14, 5], where large-scale structures emerge from local similarity relations. Literary works are thus treated as entities in a high-dimensional semantic space whose pairwise interactions give rise to observable macroscopic patterns such as communities or thematic clusters.

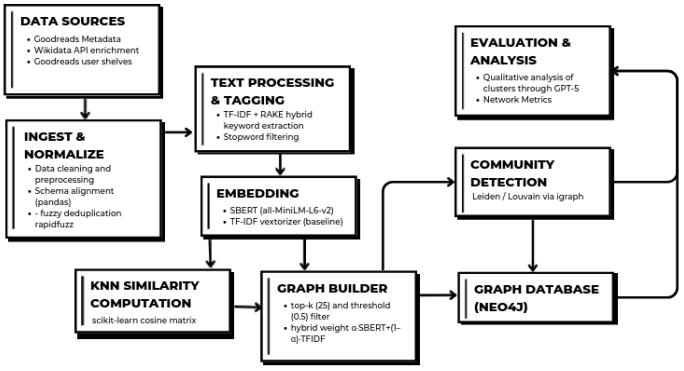


Figure 1 – Workflow of the methodological pipeline. *Source: by the author.*

2.2 DATASET DESCRIPTION AND SAMPLING

The dataset derives from the Goodreads metadata collection curated by the UCSD Recommender Systems Group [25, 26], comprising millions of bibliographic records with identifiers, titles, descriptions, and user-generated tags.

Although other sources such as Project Gutenberg provide full-text data, they lack contemporary coverage and consistent metadata, making Goodreads more suitable for large-scale relational modeling. Prior studies [22, 9] have shown that structured, semantically aligned metadata can effectively capture literary relationships at scale.

This work uses the Meta-Data of Books subset, which links works, editions, authors, series, and genre/tags. Non-thematic tags (e.g., to-read, wishlist) were removed, and missing attributes were inferred from each work's canonical "best book" edition. Sampling was performed at three levels: a 200-work pilot corpus for testing, a 5,000-work multilingual corpus for optimization, and a 10,000-work English-only corpus for scalability and community-resolution analysis. Each work includes a 100–300-word description used as the textual basis for embedding and similarity computation.

2.3 DATA CLEANING AND NORMALIZATION

Data normalization and entity resolution ensure the internal consistency of the dataset and prevent multiple representations of the same entity. In large bibliographic collections such as Goodreads, textual heterogeneity arises from variations in spelling, punctuation, and user-generated metadata. Aligning these discrepancies is essential to preserve the integrity of relationships between works, authors, and publishers.

The procedure follows three stages: metadata pruning, string normalization, and duplicate detection. Non-semantic user tags (for instance "owned", "wishlist", "currently-reading") were removed using a stop-word list combined with fuzzy matching. Remaining textual fields were normalized with *pandas* and *rapidfuzz*, applying lowercasing, punctuation stripping, and whitespace standardization.

Duplicate entities were resolved through *Levenshtein similarity*, a character-level metric that quantifies the minimum number of edits (insertions, deletions, or substitutions) required to transform one string into another [10]. This method captures typographic and orthographic variation that purely token-based measures (such as Jaccard) would miss, allowing unification of near-duplicates like "Penguin" vs "Penglin". The resulting unified dataset contains harmonized identifiers and fields prepared for enrichment , ensuring relational consistency across the dataset.

2.4 METADATA ENRICHMENT AND KEYWORD EXTRACTION

To compensate for missing metadata, the script **enrich-data.py** queried Wikidata [28] for each work and author using fuzzy-matched titles and names. Retrieved attributes included original language, country of origin, and literary movement. These supplementary fields enhance the

contextual granularity of the network, enabling later analyses such as assortativity by language or geography.

Textual enrichment relied on a hybrid keyword-extraction method combining **Term Frequency–Inverse Document Frequency (TF–IDF)** and **Rapid Automatic Keyword Extraction (RAKE)**. The rationale for merging them is that TF–IDF captures statistically relevant single terms, while RAKE identifies meaningful multi-word expressions.

The TF–ID weighting [19] quantifies how characteristic a term t is for a document d within a corpus of N documents:

$$\mathrm{tfidf}(t,d) = f_{t,d} \cdot \log \frac{N}{n_t} \tag{1}$$

where $f_{t,d}$ is the term frequency and n_t the number of documents containing t.

This function penalizes overly common words and amplifies rare, document-specific ones.

RAKE [18], on the other hand, proceeds differently. It divides the text at stop words, producing candidate phrases, then constructs a co-occurrence graph where each keyword is linked to others appearing in the same phrase. For each token t, its degree $deg(t)$ is the number of co-occurring neighbors, and its frequency $freq(t)$ counts total occurrences.

The score of a phrase p is:

$$\mathrm{score}(p) = \sum_{t \in p} \frac{\deg(t)}{\mathrm{freq}(t)} \tag{2}$$

Tokens with high co-occurrence diversity but low frequency gain prominence, highlighting semantically rich phrases such as “post-colonial identity” or “magical realism”.

After extraction, both outputs were combined: the top RAKE phrases and top 15 TF–IDF words per description were merged and deduplicated. This hybridization follows the pattern-recognition perspective in complex systems [4], capturing both global statistical significance and local contextual association.

2.5 GRAPH MODELING AND DATABASE STRUCTURE

The processed metadata was organized as a property graph, implemented in Neo4j [12], a Java-based, open-source graph database that is used to manage and query highly connected data more efficiently than traditional databases.

Neo4j’s model supports typed nodes and edges, each with their own attributes.

For example, in a film graph, nodes *Person* and *Movie* could be linked by a *WORKED_ON* relation whose property role = *"Director"*. Similarly, this project's schema (Figure 2) defines seven node types, *Work, Person, Publisher, Tag, Series, Language, and Country*, and multiple relationship types such as *WORKED_ON, PUBLISHED_BY, and PART_OF.* Each edge type includes attributes inherited from Goodreads (for example, *role* in *WORKED_ON*), preserving detailed authorship information.

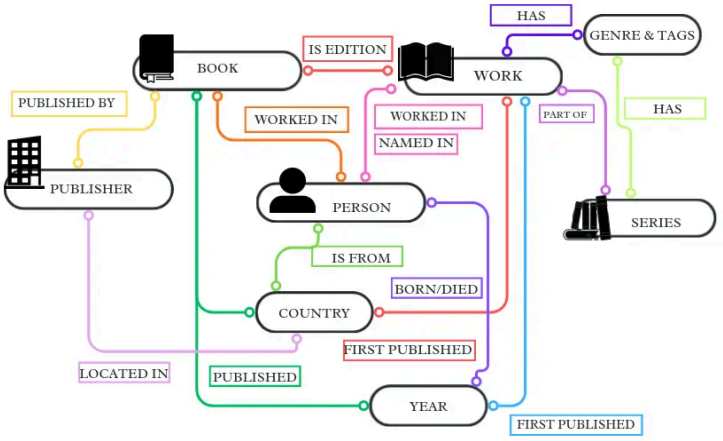


Figure 2 – Entity-relationship model of the literary knowledge graph. *Source: by the author*

2.6 EMBEDDING GENERATION AND SIMILARITY COMPUTATION

While explicit relationships (e.g., authorship or publisher) describe known links, semantic similarity between works must be inferred. This inference uses *text embeddings*, numerical vector representations capturing contextual meaning in high-dimensional space.

Each description and keyword list was encoded using **Sentence-BERT (SBERT)**, specifically the **all-MiniLM-L6-v2** model [17]. During inference, it converts any input text into a fixed-size embedding $\mathbf{v}_i \in \mathbb{R}^{384}$, this dimension corresponds to the size of the model's final hidden layer, determined during training. This enables efficient pairwise comparison via *cosine similarity*.

Two filters ensured manageable graph density:

1. Only the top $k = 25$ neighbors per node were kept;
2. Edges with $sim(i, j) < 0.55$ were discarded.

These filters serve complementary roles. The k-nearest-neighbors constraint ensures that each work remains locally connected to its most similar peers, avoiding isolated nodes and

preserving overall graph connectivity. Conversely, the similarity threshold removes weak or noisy links, improving the thematic coherence of detected communities.

Internal tests showed that using k-NN alone increased coverage but introduced low-weight edges that diluted community purity, while applying the threshold alone produced high-purity clusters at the cost of excessive fragmentation. Maintaining both filters therefore represents a practical equilibrium between structural completeness and semantic precision, yielding networks that are simultaneously interpretable and well connected.

A *hybrid weighting* incorporated both SBERT and TF–IDF similarities:

$$w_{ij} = (1 - \alpha)\,\mathrm{sim}_{\mathrm{SBERT}}(i, j) + \alpha\,\mathrm{sim}_{\mathrm{TF\text{–}IDF}}(i, j)$$

where $\alpha \in [0,1]$ controls the lexical contribution of TF–IDF. A parameter sweep varied α while monitoring graph coverage and keyword purity, identifying an optimum at α ≈ 0.11 (≈ 11 % TF–IDF weight), which balances contextual and lexical similarity. (see Section 3.2).

This weighting echoes hybrid recommender formulations where latent and explicit similarities are linearly combined [29].

2.7 COMMUNITY DETECTION

Once the weighted graph is built, *unsupervised community detection* identifies densely connected groups of works, potential analogs of genres or thematic clusters.

The *Leiden algorithm* [23] was chosen for its ability to guarantee well-connected communities. The algorithm optimizes **modularity (Q)**, a measure of how well a partition separates dense internal connections from sparse external ones: Intuitively, Q compares the observed density of intra-community edges with that expected in a random graph of identical degree distribution.

The *Leiden optimization* iterates through three phases:

1. Local movement, reassigning nodes to neighboring communities to improve modularity;
2. Refinement, splitting disconnected clusters to ensure internal connectivity;
3. Aggregation, collapsing communities into meta-nodes and repeating until Q converges.

Detected communities can be interpreted as **data-driven genre clusters**, revealing implicit literary groupings without human-labeled categories. Similar methods are used in network-based literary analysis [4, 22].

2.8 EVALUATION METRICS

The quality of the network and detected communities was evaluated using structural and semantic indicators:

1. **Graph coverage** measures the proportion of non-isolated works and indicates the connectedness of the network.
2. **Keyword purity** defined as $P = |1 - H_n|$, represents the inverse of normalized Shannon entropy. Thus, $P \approx 1$ indicates that a few keywords dominate (strongly homogeneous or thematically coherent communities), whereas $P \approx 0$ corresponds to high entropy (diverse or heterogeneous topics).

Together these metrics assess whether the graph's structure corresponds to meaningful thematic organization. A minimum of 0.30 graph coverage and 0.25 median keyword purity was adopted as a practical criterion to ensure that networks were sufficiently connected for analysis while maintaining thematic coherence within communities. These thresholds were determined empirically from pilot runs.

3 RESULTS

3.1 NETWORK OVERVIEW

The final pipeline produced two comparable literary networks derived from Goodreads metadata: a **5000-work multilingual corpus** and a **10000-work English-only corpus**.

Each work was represented as a node; similarity relations above the threshold = 0.55 and top-k = 25 were modeled as weighted undirected edges.

The complete **Neo4j property graph**, containing works, authors, publishers, series, languages, and countries, reached approximately **1.6×10^5 total relations**.

From this, the **work-to-work similarity subgraph** was extracted for analysis; because most works remained weakly connected or isolated, these graphs contained only a few thousand weighted edges, yielding a sparse but interpretable structure consistent with their reported coverage values.

Table 1 - Global network metrics

Metric	5k Works (Multi-Lingual)	10k Works (English)
Edges	*≈ 1000*	*≈ 5000*
Graph Coverage (%)	26.8	25.2
Modularity	0.904	0.865
Median Keyword Purity	0.231	0.227

Source: author's computation from Goodreads data

Both networks exhibit the sparse, heterogeneous connectivity typical of cultural and information graphs [14, 3]. The degree and weighted-degree distributions (see Figure 3) show a strong skew: a small number of highly connected works function as hubs, while most remain weakly linked. Combined with the high modularity values (≈ 0.9 and 0.86), this pattern suggests a topology consistent with small-world-like organization – locally cohesive clusters connected through a limited number of hub works. Such structure reflects the literary corpus itself, in which widely read or referenced titles bridge otherwise distinct thematic regions.

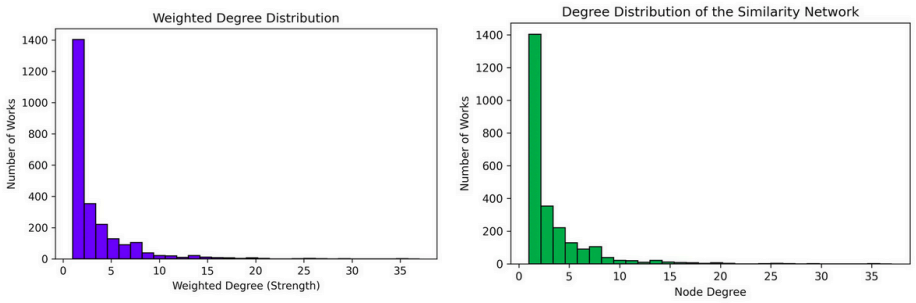
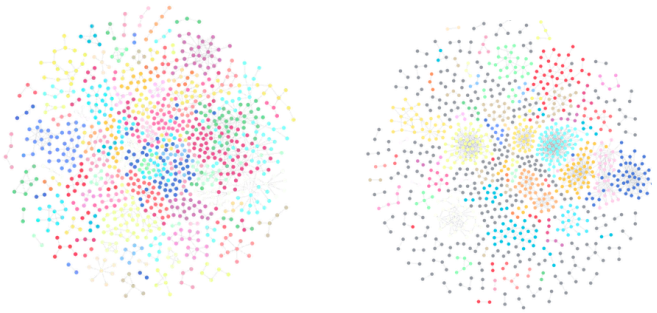


Figure 3 – Degree and weighted degree distribution showing the frequency of connections per node for the 10k node network. *Source: by the author.*

Despite the difference in corpus size, both networks show similar purity and assortativity. The multilingual dataset exhibits higher modularity because language boundaries produce clearer community divisions, while the English-only dataset achieves finer thematic resolution within a lower overall Q.



10k Works 5k Works

Figure 5 – 10k work and 5k work similarity networks with highlighted clusters *Source: by the author.*

3.2 PARAMETER OPTIMIZATION

Parameter optimization was carried out to identify the configuration that best balanced structural connectivity and thematic coherence.

A first sweep was performed with a fixed **TF–IDF weight** of $\alpha = 0.3$, serving as a baseline hybrid configuration, while varying the similarity threshold and the number of retained neighbors (top-k):

Table 2 - Threshold and Top-k Sweep ($\alpha = 0.3$)

Threshold	Top-k	Graph Coverage	Median Keyword Purity
0.5	10	0.494	0.213
0.53	10	0.339	0.308
0.53	15	0.339	0.309
0.53	20	0.297	0.309
0.54	10	0.297	0.334
0.54	15	0.257	0.334
0.55	10	0.257	0.334
0.55	15	0.257	0.334
0.55	20	0.227	0.400
0.56	10	0.227	0.400
0.56	15	0.179	0.400
0.56	20	0.179	0.462
0.58	10	0.073	0.462
0.58	20	0.073	0.636
0.65	20	0.037	0.636
0.7	25	0.037	0.667
0.7	30	0.037	0.667

Source: author's computation from Goodreads data

This initial grid provided an overview of the trade-off between network density and community cohesion. Coverage decreased as the similarity threshold increased, while median

keyword purity improved, revealing the expected inverse relationship between connectivity and thematic focus.

A second sweep then varied **α**, keeping the threshold fixed at 0.53 and top-k at 20, to evaluate the effect of the TF–IDF contribution to the hybrid similarity:

Table 3 - TF-IDF weight sweep (threshold = 0.53, top-k = 20)

Weight (α)	Graph Coverage	Median Keyword Purity
0	0.339	0.309
0.01	0.321	0.334
0.025	0.297	0.334
0.05	0.258	0.334
0.1	0.207	0.446
0.15	0.163	0.546
0.2	0.137	0.542
0.25	0.107	0.692
0.3	0.091	0.707
0.35	0.084	0.710

Source: author's computation from Goodreads data

Although α ≈ 0.01 achieved the best numeric balance between coverage and purity, this value corresponds to an almost purely SBERT-based model, undermining the purpose of a hybrid formulation.

To confirm whether a hybrid component remained beneficial, a third sweep was conducted with a stricter similarity threshold (0.55) to ensure a minimum semantic relation between works, and top-k = 25 to compensate for reduced density:

Table 4 - TF-IDF weight sweep (threshold = 0.55, top-k = 25)

Weight (α)	Graph Coverage	Median Keyword Purity
0	0.494	0.220
0.025	0.445	0.289
0.08	0.334	0.334
0.09	0.316	0.334
0.1	0.301	0.346
0.12	0.272	0.334
0.15	0.238	0.384
0.18	0.209	0.435
0.2	0.192	0.462
0.3	0.133	0.696

Source: author's computation from Goodreads data

The results demonstrate a stable improvement in thematic cohesion up to α ≈ 0.10–0.12, beyond which purity gains plateau while coverage declines sharply.

Accordingly, the final configuration adopted for subsequent analyses was α = 0.10, threshold = 0.55, and top-k = 25.

This setting preserves a modest lexical contribution (~10 % TF–IDF) that enhances community coherence without fragmenting the network, confirming the advantage of a hybrid similarity representation.

To validate the parameter selection, the data from Table 4 was plotted as a relation between graph coverage and median keyword purity (Figure 6). The resulting curve exhibits an exponential decay, indicating the expected trade-off between connectivity and thematic cohesion. An exponential model of the form $y = ae^{-bx} + c$ was fitted to the data, and a piecewise linear approximation was applied to the fitted curve to locate the curvature-based elbow point, corresponding to the intersection of two locally linear regions.

This analysis identified an optimal TF–IDF weight of α ≈ 0.125 (Figure 6, red marker). However, since the measured purity values show a slight decline beyond α ≈ 0.12 (see Table 4), the configuration α = 0.10, threshold = 0.55, and top-k = 25 was retained as the final operational setting. This choice maintains maximum stability while preserving strong thematic coherence without excessive network fragmentation.

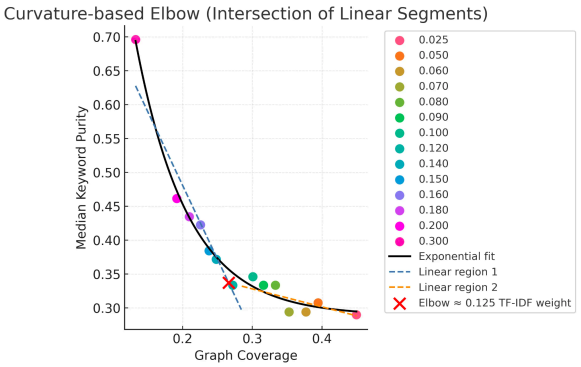


Figure 6 – Relation between median keyword purity and graph coverage for different thresholds.
Source: by the author.

3.3 QUALITATIVE CLUSTER ANALYSIS

After community detection, each network was decomposed into clusters corresponding to emergent thematic or stylistic groups. To assign interpretive labels, a **hybrid human–AI procedure** was employed: the author manually verified and refined preliminary categorizations generated by OpenAI's GPT-5 model, which analyzed top-ranked keywords for each community. This assisted interpretation ensured consistency while maintaining human oversight, in accordance with the methodological framework stated in the project proposal.

3.3.1 Multilingual network

The 5000-work graph produced 71 communities, 17 of which contained more than 10 works. Because the dataset included multiple languages, clusters tended to **align with linguistic or regional boundaries** rather than purely thematic affinities.

One large cluster composed of tokens that were mis-codified inside of the database, as they were not alpha-numeric, seeming at first like random choice letters (Figure 7, Cluster 2) proved to be **Arabic-script text** when properly interpreted by utf-8, its encoding was not recognized by the English stop-word filter, resulting in a distinct "language artefact" cluster. Other large clusters grouped together Northern-European literary fiction, Russian-language works, and Latin-American poetry.

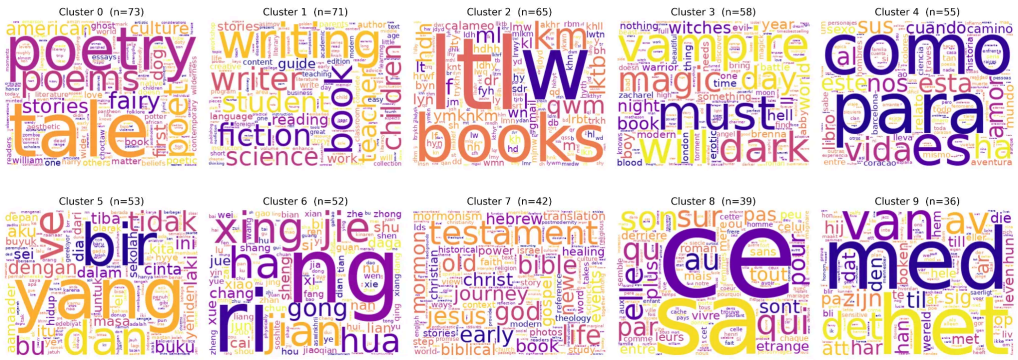


Figure 7 – Word clouds representing the top keywords for the largest clusters in the 5k work network. *Source: by the author.*

3.3.2 English-only network

When restricted to English texts, the network produced 90 communities, 22 with more than 20 works. Linguistic uniformity removed the language-based separation seen earlier, allowing the algorithm to discriminate **subgenres within single literary domains**. Whereas the multilingual graph is clustered by region, the English-only graph is clustered by narrative content and tone.

This exemplifies the improved granularity achieved through linguistic normalization: the model differentiates overlapping but distinct universes based purely on textual metadata.

Figure 8 – Word clouds representing the top keywords for the largest clusters in the 10k work network.
Source: by the author.

3.4 LANGUAGE ANALYSIS

3.4.1 Language metadata and encoding artifacts

Although the enrichment pipeline queried Wikidata for language metadata, few records contained valid or consistent entries, and many used generic placeholders such as “English (United States).” Consequently, language purity could not be reliably quantified.

All preprocessing steps – tokenization, lemmatization, and stop-word removal – used primarily English filters. Attempts to include comprehensive multilingual stop-word lists proved impractical. Descriptions written in other languages or non-Latin scripts were therefore not properly tokenized, causing every token to be treated as valid and inflating pairwise similarity among those works. The resulting language-based clusters, while artefactual, illustrate the model’s sensitivity to systematic inconsistencies in the data.

Table 5 – Most connected cluster in the 5k Work Network (Arabic Encoding Artefact, average degree = 14.2)

Median Keyword Purity	Dominant Language	Top Keywords (freq > 5)	Representative Titles (sample)
0.89	Arabic (script)	كاتب, رواية, الأدب, الحب, القاهرة, الحياة	*قصائد مختارة* (Selected Poems), *رواية الحياة* (The Novel of Life), *أدب القاهرة* (Cairo Literature)

Source: identified by the author during qualitative keyword analysis with GPT-5

A strong example is the identified Arabic works cluster. Even though this community results from incorrect encodings rather than semantics, its isolation by the algorithm is methodologically relevant. It shows that the similarity model and community detection pipeline are capable of flagging coherent patterns that deviate sharply from the rest of the network, an important property for quality control in large heterogeneous datasets. Such behaviour can be exploited in future work for anomaly detection, helping to identify clusters formed by metadata inconsistencies, duplicated records, or language misclassifications.

3.5 NETWORK METRICS AND SUCCESS EVALUATION

To assess the validity of the constructed networks, quantitative indicators from both runs were compared (Table 6). All metrics were computed using deterministic pipelines, ensuring reproducibility across iterations.

Table 6 - Global network metrics

Network	Graph Coverage (%)	Modularity (Q)	Median Keyword Purity
5k Works (Multilingual)	26.8	0.904	0.231
10k Works (English only)	25.2	0.865	0.227

Source: author's computation from Goodreads data

Both networks display consistent modularity and purity, indicating that doubling the corpus size did not compromise community quality. The 10 k-work graph shows a slight reduction in modularity – a natural consequence of increased inter-cluster overlap among English texts, but its larger size enhances genre resolution, revealing more finely separated subgenres. Tag purity,

computed from Goodreads user shelves, remained close to zero, confirming that community coherence arises from textual similarity rather than social tagging behavior. This divergence underscores the advantage of semantic-embedding approaches for cultural network modeling [4].

Overall, both the statistical and qualitative results demonstrate that the pipeline yields stable, high-modularity literary networks with interpretable thematic clusters, validating the hybrid embedding and parameter-optimization methodology.

4 CONCLUSIONS AND FINAL CONSIDERATIONS

This work modeled literary similarity as a complex network, showing how large bibliographic datasets can uncover emergent thematic and stylistic patterns. Using Goodreads metadata enriched via Wikidata, two semantic networks were constructed: one with 5,000 multilingual works and another with 10,000 English-language works, linked through hybrid similarity combining lexical and contextual embeddings.

The framework integrated TF–IDF and RAKE keyword extraction, Sentence-BERT embeddings, and Leiden community detection in a reproducible Python–Neo4j pipeline. Parameter sweeps and exponential fitting identified an optimal configuration (threshold = 0.55, top-k = 25, α = 0.11) balancing connectivity and coherence. The system exhibited transition-like behavior typical of complex systems: as thresholds rose, the graph fragmented; as they fell, thematic boundaries blurred. The resulting networks showed high modularity ($Q \approx 0.9$) and consistent keyword purity ($\approx$ 0.23), indicating stable, interpretable communities that reflect meaningful literary relations.

Although overall coverage remained low due to sparse textual overlap, identified clusters aligned with recognizable literary domains—poetry, fiction, science fiction, romance, and comics—supporting the hypothesis that genres emerge as topological communities in semantic space. Comparing both corpora revealed that linguistic uniformity enhances thematic resolution: while the multilingual network clustered by language and region, the English-only graph resolved fine-grained subgenres such as young-adult romance or distinct comic-book universes.

The results also validated a hybrid human–AI interpretation process, where GPT-assisted labeling complemented manual verification to maintain both scale and accuracy. Remaining limitations include incomplete Goodreads metadata and the reliance on descriptions rather than full texts, which restricts semantic depth. Nonetheless, the study demonstrates that hybrid embedding

and network-based modeling can capture interpretable cultural structures from large-scale bibliographic data.

4.1 FUTURE DIRECTIONS

Future extensions of this work may include:

1. Full-text embeddings — Applying document-level models to complete literary works to capture narrative structure and stylistic depth.
2. Temporal evolution modeling — Building dynamic graphs to trace the propagation of genres and stylistic features over time.
3. Cross-lingual integration — Using multilingual embedding models to bridge language barriers and align regional corpora within a unified semantic space.
4. Interactive visualization platform — Implementing an exploration interface in Dash + Cytoscape or Neo4j Bloom, allowing users to navigate clusters and inspect relationships visually.
5. Recommendation and influence systems — Leveraging the hybrid similarity model to create tools for suggesting related works or mapping literary influence.

Ultimately, this study confirms that the methods of complex systems and machine learning can provide quantitative insight into cultural structures. Even using limited textual information, the resulting networks reproduce recognizable literary groupings and uncover latent relations invisible to traditional cataloguing systems. This convergence of data science and literary analysis points toward a future where knowledge graphs and embeddings serve not merely as technical artifacts but as instruments for expanding how we perceive and navigate the world's written culture.

REFERENCES

[1] AFOUDI, Y.; LAZAAR, M.; ACHHAB, M. A. **Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network.** *Simulation Modelling Practice and Theory*, v. 113, p. 102375, 2021. DOI: https://doi.org/10.1016/j.simpat.2021.102375.

[2] BARABÁSI, A.-L. **Scale-free networks: A decade and beyond.** *Science*, v. 325, n. 5939, p. 412–413, 2009. DOI: https://doi.org/10.1126/science.1173299.

[3] BARABÁSI, A.-L.; ALBERT, R. **Emergence of scaling in random networks.** *Science*, Washington, v. 286, n. 5439, p. 509-512, 1999. DOI: https://doi.org/10.1126/science.286.5439.509.

[4] COSTA, L. da F. *et al.* **A pattern recognition approach to complex networks.** *Journal of Statistical Mechanics: Theory and Experiment*, v. 2010, n. 11, p. P11015, 2010. DOI: https://doi.org/10.1088/1742-5468/2010/11/P11015.

[5] FORTUNATO, S.; HRIC, D. **Community detection in networks: a user guide.** *Physics Reports*, v. 659, p. 1-44, 2016. DOI: https://doi.org/10.1016/j.physrep.2016.09.002.

[6] GOODREADS. **Goodreads: book discovery and community platform.** Available at: https://www.goodreads.com/. Accessed on: 5 Nov. 2025.

[7] GOODREADS DATASET. **Goodreads datasets for research.** Curated by the UCSD Recommender Systems Group. Available at: https://cseweb.ucsd.edu/~jmcauley/datasets/goodreads.html. Accessed on: 5 Nov. 2025.

[8] IGRAPH PROJECT. **igraph Library for Complex Network Analysis.** Available at: https://igraph.org/. Accessed on: 5 Nov. 2025.

[9] KOKASH, I. *et al.* **From books to knowledge graphs.** *Data Intelligence*, v. 5, n. 2, p. 329-357, 2023. DOI: https://doi.org/10.1162/dint_a_00165.

[10] LEVENSHETEIN, V. I. **Binary codes capable of correcting deletions, insertions, and reversals.** *Soviet Physics Doklady*, Moscow, v. 10, n. 8, p. 707-710, 1966.

[11] MATPLOTLIB DEVELOPMENT TEAM. **Matplotlib: visualization library for Python.** Available at: https://matplotlib.org/. Accessed on: 5 Nov. 2025.

[12] NEO4J. **Neo4j Graph Database Platform.** Neo4j Inc. Available at: https://neo4j.com/. Accessed on: 5 Nov. 2025.

[13] NETWORKX PROJECT. **NetworkX: graph analysis in Python.** Available at: https://networkx.org/. Accessed on: 5 Nov. 2025.

[14] NEWMAN, M. E. J. **The structure and function of complex networks.** *SIAM Review*, v. 45, n. 2, p. 167-256, 2003. DOI: https://doi.org/10.1137/S003614450342480.

[15] NUMPY DEVELOPERS. **NumPy: fundamental package for scientific computing with Python.** Available at: https://numpy.org/. Accessed on: 5 Nov. 2025.

[16] PANDAS DEVELOPMENT TEAM. **pandas: data analysis library for Python.** Available at: https://pandas.pydata.org/. Accessed on: 5 Nov. 2025.

[17] REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence embeddings using Siamese BERT-networks.** In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Hong Kong: Association for Computational Linguistics, 2019. p. 3982-3992. DOI: https://doi.org/10.18653/v1/D19-1410.

[18] ROSE, S. *et al.* **Automatic keyword extraction from individual documents.** In: BERRY, M. W.; KOGAN, J. (orgs.). *Text Mining: Applications and Theory.* Chichester: John Wiley & Sons, 2010. p. 1-20. DOI: https://doi.org/10.1002/9780470689646.ch1.

[19] SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval.** New York: McGraw-Hill, 1983.

[20] SCIKIT-LEARN DEVELOPERS. **scikit-learn: machine learning in Python.** Available at: https://scikit-learn.org/. Accessed on: 5 Nov. 2025.

[21] SENTENCE-TRANSFORMERS PROJECT. **Sentence-Transformers: state-of-the-art sentence and text embeddings.** Available at: https://www.sbert.net/. Accessed on: 5 Nov. 2025.

[22] STRANISCI, M. A. *et al.* **The World Literature Knowledge Graph.** *Proceedings of the 22nd International Semantic Web Conference (ISWC 2023)*, Athens, Greece, p. 1-15, 2023. Available at: https://literaturegraph.di.unito.it/. Accessed on: 5 Nov. 2025.

[23] TRAAG, V. A.; WALTMAN, L.; VAN ECK, N. J. **From Louvain to Leiden: guaranteeing well-connected communities.** *Scientific Reports*, London, v. 9, n. 1, p. 5233, 2019. DOI: https://doi.org/10.1038/s41598-019-41695-6.

[24] WAN, M.; MISRA, R.; NAKASHOLE, N.; MCAULEY, J. **Fine-grained spoiler detection from large-scale review corpora.** In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Florence: Association for Computational Linguistics, 2019. p. 3741-3751. DOI: https://doi.org/10.18653/v1/P19-1368.

[25] WAN, M.; MCAULEY, J. **Item recommendation on monotonic behavior chains.** In: *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys 2018)*. Vancouver: ACM, 2018. p. 86-94. DOI: https://doi.org/10.1145/3240323.3240375.

[26] WAN, M.; MCAULEY, J. **Learning to embed categorical features without embedding tables for recommendation.** In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Florence: Association for Computational Linguistics, 2019. p. 872-884. DOI: https://doi.org/10.18653/v1/P19-1084.

[27] WATTS, D. J.; STROGATZ, S. H. **Collective dynamics of "small-world" networks.** *Nature*, v. 393, p. 440-442, 1998. DOI: https://doi.org/10.1038/30918.

[28] WIKIDATA. **Wikidata: a free and open knowledge base.** Wikimedia Foundation. Available at: https://www.wikidata.org/. Accessed on: 5 Nov. 2025.

[29] ZHANG, Y.; ZHANG, J.; LI, M. **A novel hybrid deep recommendation system to differentiate user's preference and item's attractiveness.** *Neural Networks*, v. 130, p. 1-14, 2020. DOI: https://doi.org/10.1016/j.neunet.2020.07.002.