UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Previsão de padrões de sucesso de livros mediante o uso de redes complexas

Maryory Loaiza Agudelo

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

Maryory Loaiza Agudelo

Previsão de padrões de sucesso de livros mediante o uso de redes complexas

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadores:
Prof. Dr. Diego Raphael Amancio
Prof. Dra. Solange Oliveira Rezende

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Loaiza-Agudelo, Maryory
L795p Previsão de padrões de sucesso de livros mediante o uso de redes complexas / Maryory Loaiza-Agudelo; orientador Diego Raphael Amancio Amancio; coorientadora Solange Rezende. -- São Carlos, 2023.
82 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. . I. Amancio, Diego Raphael Amancio, orient. II. Rezende, Solange, coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

À minha amada filha Sabina.

AGRADECIMENTOS

A oportunidade de fazer o MBA em Big Data e Inteligência Artificial chegou na minha vida para dar o plus que eu tanto estava procurando para minha carreira profissional. Assim que só posso agradecer:

- Á coordenação do programa pela oportunidade de fazer parte dessa turma e pelo apoio recebido durante este periodo,
- Ao meus orientadores pelo acompanhamento,
- A Xiomara Quispe pelos ensinamentos,
- A todos os professores, pesquisadores e tutores, que com seu compromiso e conhecimento, fizeram do aprendizado um processo construtivo e de grande crescimento,
- A minha familia pelo apoio constante e por acreditarem sempre em mim, fornecendo o amor e os espaços de estudo,
- A Universidade de São Paulo, pelos programas de qualidade que oferece à comunidade.

"La vida no es fácil, para ninguno de nosotros.
Pero... ¡qué importa! Hay que perseverar y, sobre todo, tener confianza en uno mismo.
Hay que sentirse dotado para realizar alguna cosa y que esa cosa hay que alcanzarla,
cueste lo que cueste".
Marie Curie

ABSTRACT

Loaiza-Agudelo, M **Prediction of book success patterns using complex networks**. 2023. 82p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Reading books is an important factor in the modern world, it is part of human culture and education and is a favorite activity for many people. Which makes the bookselling industry big business these days. However, the sales success of a given book depends on factors such as: the writing style, critics, book reviews, awards received, advertising, publicity on social networks, as well as comments from readers themselves.

Despite the fact that several researchers have studied the characteristics that influence the success of books, and currently research groups have focused their efforts on determining these characteristics, in general, the prediction of this success based on different factors specific to books has received little attention. And the advance prediction of success of a book of a given genre, in addition to making it possible to predict the profits generated by sales, provides information about readers' preferences regarding authors and plots, the means of publication (on paper or online), the means of promoting books, as well as market needs for future publications. But predicting the success of a book or literary genre before its release and understanding the mechanisms behind its success or failure is not a simple task. Currently, the publishing industry provides limited information to editors to help them make decisions about publishing, such as: how many copies to print, investment in marketing, launch date. Therefore, publishers base their decision on factors such as the authors' previous success, rather than relying on data specifically linked to the book being considered for publication. In this way, the advance prediction of the book's success through pre-publication information can be fundamental for making decisions before publication.

In order to offer better predictions of a book's success, our efforts are focused on determining patterns, exploring factors such as connections between characters. This process involves natural language processing (NLP) in pre-processing the data, -obtained from a list composed of novels by different authors, retrieved from the Project Gutenberg dataset-, complex networks in creating the network, and machine learning in analyzing the social networks (relationships between characters).

Keywords: complex networks; anaphora; machine learning; tokenization; book success.

RESUMO

Loaiza-Agudelo, M **Previsão de padrões de sucesso de livros mediante o uso de redes complexas**. 2023. 82p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A leitura de livros é um fator importante no mundo moderno, faz parte da cultura e a educação do ser humano e, é a atividade favorita para muitas pessoas. O que faz que a industria de venda de livros seja um grande negócio atualmente. Mas, o sucesso das vendas de determinado livro depende de fatores como: o estilo de escrita, os críticos, as resenhas dos livros, os prêmios recebidos, os anúncios publicitários, a divulgação nas redes sociais, assim como os comentários dos próprios leitores.

Apesar do fato de que vários pesquisadores têm estudado as características que influenciam no sucesso dos livros, e atualmente grupos de pesquisa têm focado seus esforços na determinação destas características, em termos gerais, a previsão deste sucesso a partir de diferentes fatores próprios dos livros tem recebido pouca atenção. E a previsão antecipada de sucesso de um livro de determinado gênero, além de permitir prever os lucros gerados pelas vendas, fornece informações sobre as preferencias dos leitores enquanto a autores e tramas, os meios de publicação (no papel ou on-line), os meios de promoção dos livros, assim como as necessidades do mercado para futuras publicações. Mas, prever o sucesso de um livro ou um gênero literário antes de seu lançamento e compreender os mecanismos por trás de seu sucesso ou fracasso não é uma tarefa simples. Atualmente a indústria editorial disponibiliza informações limitadas aos editores, que permitam auxiliar em suas decisões sobre a publicação, como: quantas cópias imprimir, investimento em marketing, data de lançamento. Pelo que as editoras baseiam a sua decisão em fatores como o sucesso anterior dos autores, em vez de confiar em dados especificamente vinculados ao livro considerado para publicação. Dessa maneira, a previsão antecipada do sucesso do livro mediante informações de pré-publicação, pode ser fundamental para a toma de decisões antes da publicação.

No intuito de oferecer melhores previsões do sucesso de um livro, nossos esforços são direcionados na determinação de padrões, explorando fatores como as conexões entre personagens. Este processo envolve processamento de linguagem natural (PLN) no pré-processamento dos dados, -obtidos de uma lista composta de romances de autores distintos, recuperados do *Project Gutenberg dataset*, redes complexas na criação da rede, e aprendizado de máquina na análise das redes sociais (relacionamento entre as personagens).

Palavras-chave: sucesso de livros; redes complexas; anafora; aprendizado de máquina; tokenização.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

KNN	K-Nearest-Neighbor
NE	Named Entity
NER	Named Entity Recognition
TNYTBS	The New York Times Best Sellers
MLP	Multilayer Perceptron
RFO	Random Forest
SVM	Support Vector Machines
PWBL	Publishers Weekly Bestseller Lists

SUMÁRIO

1 INTRODUÇÃO

A leitura de livros é um fator importante no mundo moderno, faz parte da cultura e a educação do ser humano e, é a atividade favorita para muitas pessoas. O que faz que a industria de venda de livros seja um grande negócio atualmente. Mas, o sucesso das vendas de determinado livro depende de fatores como: o estilo de escrita, os críticos, as resenhas dos livros, os prêmios recebidos, os anúncios publicitários, a divulgação nas redes sociais, assim como os comentários dos próprios leitores (Wang *et al.*, 2019). Particularmente, Yucesoy *et al.* (2018), estudaram o sucesso de livros a partir de uma amostra de *bestsellers* publicada em "*the New York Times bestseller list (NYTBL)*", encontrando que este tipo de livros têm maior chance de pertencer às categorias gerais de ficção e biografias e, independentemente do subgênero, os livros de não ficção vendem menos cópias do que os livros de ficção. Mas, esta pesquisa é limitada a apenas dois gêneros de livros, e o mundo literário tem uma ampla variedade de gêneros.

No entanto, apesar do fato de que vários pesquisadores têm estudado as características que influenciam no sucesso dos livros (Clement; Proppe; Rott, 2007; Beck, 2007; Schmidt-Stölting; Blömeke; Clement, 2011; Nakamura, 2013; Shehu *et al.*, 2014; Yucesoy *et al.*, 2018), e atualmente grupos de pesquisa têm focado seus esforços na determinação destas características, em termos gerais, a previsão deste sucesso a partir de diferentes fatores próprios dos livros tem recebido pouca atenção. Segundo Wang *et al.* (2019), um estudo publicado nesta área (Schmidt-Stölting; Blömeke; Clement, 2011), focado na venda de livros no mercado alemão, aplicou um modelo linear, relatando uma precisão limitada. Por outro lado, a previsão antecipada de sucesso de um livro de determinado género, além de permitir prever os lucros gerados pelas vendas, fornece informações sobre as preferências dos leitores enquanto a autores e tramas, os meios de publicação (no papel ou on-line), os meios de promoção dos livros, assim como as necessidades do mercado para futuras publicações. Contudo, prever o sucesso de um livro ou um género literario antes de seu lançamento e compreender os mecanismos por trás de seu sucesso ou fracasso não é uma tarefa simples.

1.1 Justificativa e motivação

Atualmente a indústria editorial disponibiliza informações aos editores para ajudá-los em suas decisões sobre a publicação, como: quantas cópias imprimir, investimento em marketin e data de lançamento, mas estas informações são muito limitadas; pelo que as editoras baseiam a sua decisão em fatores como o sucesso anterior dos autores, em vez de confiar em dados especificamente vinculados ao livro considerado para publicação.

Dessa maneira, a previsão antecipada do sucesso do livro mediante informações de pré-publicação, pode ser fundamental para a toma de decisões antes da publicação. No intuito de oferecer melhores previsões do sucesso de um livro, esta pesquisa é direccionada na determinação de padrões através de redes complexas, explorando fatores como as conexões entre personagens.

1.2 Objetivos

O objetivo principal desta pesquisa é prever o sucesso de livros, implementando redes complexas na determinação dos padrões relevantes, fornecidos principalmente pelas conexões entre personagens. Por tanto, devemos considerar quais são as características das histórias que favorecem a previsão de sucesso, assim como se a sintaxe ou semântica são importantes nesta previsão. Levando ao desenvolvimento de atributos que podem caracterizar os livros, classificando-os entre sucesso vs. não sucesso. Dessa maneira, foram definidos os seguintes objetivos de trabalho:

- Criação de algoritmos que permitam identificar os diferentes padrões relacionados com o sucesso de um livro.
- Investigação quantitativa das características que favorecem a previsão de sucesso, por exemplo, se ficar voltando a um mesma "cena" ou se uma historia linear favorece mais ou menos o sucesso.
- Implementação de redes de co-ocorrência na determinação de padrões sintáticos e semânticos em divisões dos livros, a fim de obter previsões de sucesso.

Esperando que os resultados levem a inovar nos padrões que devem ser considerados na pre-publicação de livros, que garantam um maior sucesso de aceitação e vendas.

2 FUNDAMENTAÇÃO TEÓRICA

Quando falamos em redes complexas, é necessário compreender primeiro o que são os **grafos**. Estes são objetos matemáticos constituidos por um conjunto de **nós** (chamados de vértices) conectados mediante uniões (arestas), formando uma estrutura mais ou menos complexa. Existem na natureza inúmeros sistemas cujo comportamento emergente é o resultado da interação entre um grande número de partes, as quais interagem entre si através de diversos fatores, sendo que sua estrutura pode ser definida em termos de uma rede complicada de nós e uniões entre elas, onde cada parte é definida por uma váriavel de estado associada ao nó, e onde o estado das uniões define a influência ou interação entre os nós.
A partir das propriedades desta rede de estrutura complexa, pode-se inferir certo comportamento emergente no sistema. Por tanto, podemos disser que os grafos permitem codificar relacionamentos (por exemplo: amizade) entre pares de objetos (pessoas, cidades, empresas, filmes, livros, etc.).

2.1 Redes complexas

As redes complexas também são conhecidas como grafos, são sistemas complexos que representam muitas partes que interagem entre si. Estas redes possuem então, uma coleção de **vértices** e **arestas**. Na figura 1 é apresentado um exemplo de rede circular.

O estudo de redes complexas tem sido coberto por um ramo da matemática discreta chamado **"teoria dos grafos"**. A seguir alguns conceitos importantes:

1. Um **grafo não direcionado** $G = (\mathcal{N}, \mathcal{L})$, é composto de dois conjuntos, $\mathcal{N} = \{n_1, ..., n_N\}$ cujos elementos são os **nós, vértices ou pontos do grafo**, e $\mathcal{L} = \{l_1, ..., l_L\}$ cujos elementos são as **uniões, arestas ou linhas do grafo**. Por tanto, um grafo possui N nós e L arestas e é denotado como $G(N,L)$ ou simplesmente $(\mathcal{N}, \mathcal{L})$.

2. Cada vértice ou nó n_i é determinado por sua ordem i no conjunto $\mathcal{N}$.

3. As arestas são definidas pelas ordens dos nós que elas unem, ou seja, a união entre os vértices n_i e n_j é denotada por $l_k = (i, j) = (n_i, n_j) = l_{ij}$. Neste caso, se houver união ou aresta entre dois nós, os nós são chamados de **vizinhos** ou **adjacentes**.

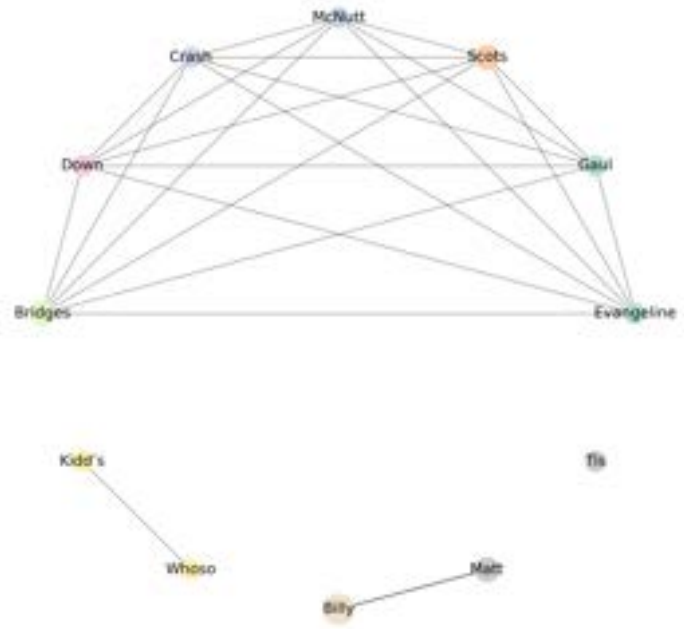

Figura 1 – Exemplo de rede circular, construida como teste para um dos livros das nossas amostras, e baseada no relacionamento entre as personagens do livro.

4. Se uma aresta é da forma $l_k = (i, i)$, isto é, ela conecta o nó i consigo mesmo, isto é chamado de ***loop*, auto-juntivo ou auto-conectável (*tadpole*)**. Estes podem aparecer em redes direcionadas e não direcionadas.

5. Duas arestas que unem dois mesmos vértices são chamadas **paralelas**.

6. As arestas que possuem um vértice em comum são chamadas de **arestas adjacentes**.

2.1.1 Tipos de grafos

1. **Grafo não direcionado:** é aquele em que a ordem dos índices nas uniões é o mesmo ($l_{ij} = l_{ji}$).

2. **Grafo direcionado:** a ordem dos índices nas uniões é improtante para que $l_{ij} \neq l_{ji}$.

3. **Multigrafos:** são grafos que possuem autouniões ou *loops*, por exemplo l_{ii}, ou multiplas uniões entre os mesmos dois nós. Os grafos não direcionados que não possuem *loops* ou arestas paralelas são chamados de **grafos simples**.

4. **Grafos pesados:** são grafos nos quais cada aresta recebe um peso ou valor numérico que mede a intensidade da união. Caso contrário, a rede ou grafo é chamado de **não pesado**. Se todos os valores tiverem o mesmo sinal (positivo ou negativo) a rede é

dita ***"unsigned"***, e se os pesos das arestas tiverem sinais diferentes associados a elas, então o grafo é ***"signed"***.

5. **Grafo vazio:** é um grafo sem arestas, apenas com nós.

6. **Grafo nulo:** é um grafo que não possui vértices, e por tanto não possui arestas.

7. **Numero possível de uniões:** para um grafo não direcionado *G(N,L)*, o número possível de uniões L está entre 0 e $N(N - 1)/2$.

8. Um grafo é dito diluído ou escasso se $L \ll N^2$, e denso se $L \sim O(N^2)$

9. Um *N-grafo* **completo** $G = (N, L)$, é aquele tal que $L = N(N - 1)/2$, e é denotado por L_N. Em um grafo completo, cada par de nós é conectado por uma única aresta. Um grafo L_3 é chamado de **triângulo**. Um grafo completo é um grafo regular com todos os seus vértices de grau $N - 1$.

10. **Árvores:** na teoria dos grafos, uma árvore é um grafo não direcionado no qual qualquer par de nós é conectado exatamente por uma única união. Um conjunto disjunto de árvores é uma floresta. Uma árvore na qual um nó está conectado a todos os outros é chamada de **rede estrela**, conforme ilustrado na Figura 2.

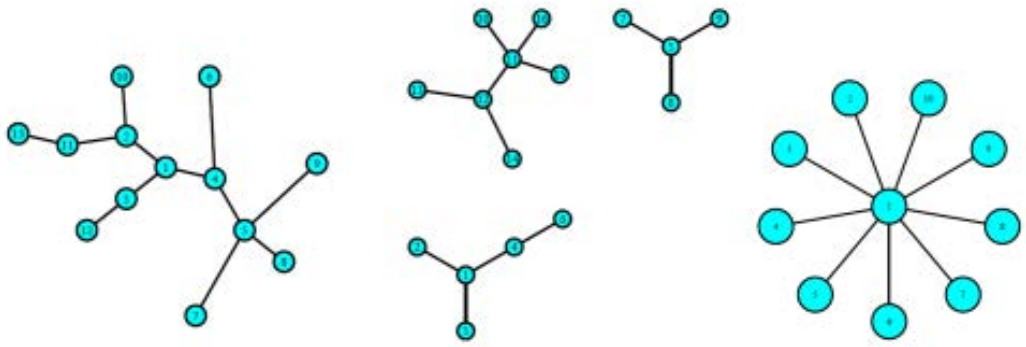

Figura 2 – Exemplos de árvore, floresta e estrela (Universidad de Granada,).

2.1.2 Representação gráfica de grafos

Os grafos são representados graficamente por redes onde os nós são representados por pontos e as uniões por linhas que conectam pontos adjacentes (ver Figura 3). Para caracterizar essas redes são utilizados dois tipos de representações, a **lista de arestas** e a **matriz de adjacência**. Ambas as caracterizações dependerão do tipo de grafo que estamos estudando.

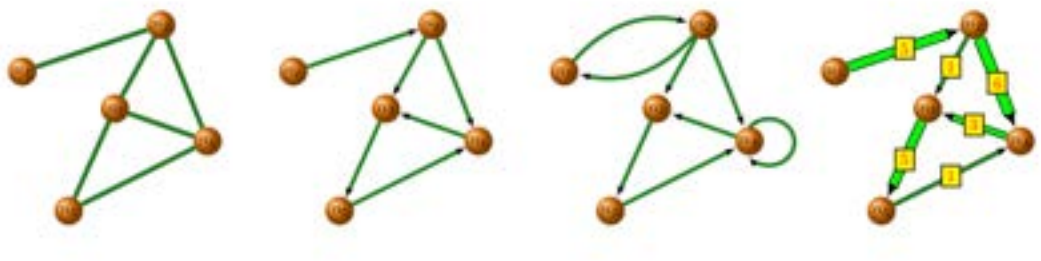

Figura 3 – Representação de alguns tipos de grafos. Da esquerda para a direita: grafo não direcionado (também é um grafo simples), grafo direcionado, multigrafo, grafo pesado (Universidad de Granada,).

2.1.2.1 Lista de arestas e matriz de adjacência de um grafo simples

1. **Lista de arestas:** No caso de um grafo simples, a lista de arestas é uma lista de $\mathcal{L}$ pares de nós (n_j, n_i), indicando que o nó j está unido a i por uma aresta. Para um grafo simples, a lista de arestas não admite redundâncias, ou seja, se o par (j, i) existir, o par (i, j) não está incluído nela.

2. **Matriz de adjacência:** a matriz de adjacência A de um grafo simples de $\mathcal{N}$ nós é uma matriz quadrada $N \times N$ cujos elementos são:

$$A_{i,j} = 1 \quad \text{se o nó } j \text{ está unido a } i \text{ por uma aresta.} \tag{2.1}$$

$$A_{i,j} = 0 \quad \text{para qualquer outra situação.} \tag{2.2}$$

2.1.2.2 Lista de arestas e matriz de adjacência de um grafo dirigido

1. **Lista de arestas:** neste caso é uma lista formada por $\mathcal{L}$ pares de nós ordenados (n_j , n_i) = (j, i) indicando que o nó j aponta para o nó i na aresta direcionada. $\mathcal{L}$ indica, portanto, o número de arestas direcionadas na rede.

2. **Matriz de adjacência:** a matriz de adjacência A é definida para um grafo direcionado como a matriz quadrada $N \times N$ com elementos:

$$A_{i,j} = 1 \quad \text{se o nó } j \text{ aponta ao nó } i. \tag{2.3}$$

$$A_{i,j} = 0 \quad \text{para qualquer outra situação.} \tag{2.4}$$

2.1.3 Medidas de redes

As propriedades estruturais mais fundamentais de uma rede complexa são a quantidade de nós e a quantidade de uniões. Na maioria dos casos estes números são grandes, portanto a análise de redes complexas requer o uso de técnicas computacionais e o uso do computador.

Neste projeto foram usadas algumas medidas de centralidade importantes para a análise do relacionamento entre as personagens das amostras de livros, tais como: grau, intermediação (*Betweenness*), proximidade (*Closeness*), coeficiente de *clustering*, entre outras. Por isto, a seguir estas são definidas:

1. **Número total de arestas:** em uma rede complexa pode ser calculado diretamente a partir da matriz de adjacência.

   – Rede não direcionada com autoconexões:

   $$L = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii}. \tag{2.5}$$

   – Rede não direcionada sem autoconexões:

   $$L = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}. \tag{2.6}$$

   – Rede direcionada:

   $$L = \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}. \tag{2.7}$$

2. **Grau de um nó:** o grau das redes complexas é definido pelo número de arestas conetadas ao vértice, ou seja, é o número de vizinhos que um vértice tem. Em uma rede não direcionada se define o **grau de um nó** i como o número total de arestas incidentes no referido nó e é denotado por k_i. Em uma rede direcionada, é feita uma distinção entre o **grau de entrada** de um nó i como o número total de nós que apontam para esse nó e é representado por k_i^{in}, e o **grau de saída** de um nó i como o número total de nós apontados pelo nó i e denotados por k_i^{out}. No caso de redes não pesadas, o grau de um nó pode ser calculado diretamente a partir da matriz de adjacência:

   $$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{j=1}^{N} A_{ji} \quad \text{rede não direcionada.} \tag{2.8}$$

$$k_i^{in} = \sum_{j=1}^{N} A_{ij} \quad \text{rede direcionada.} \tag{2.9}$$

$$k_i^{out} = \sum_{j=1}^{N} A_{ji} \quad \text{rede direcionada.} \tag{2.10}$$

Em um grafo simples não direcionado $k_i \in [0, N-1]$, e em uma rede direcionada $k_i^{in} \in [0, N-1]$ e $k_i^{out} \in [0, N-1]$

3. **Sequência de graus:** em uma rede não direcionada, falamos em sequência de graus como a sequência ordenada dos graus de todos os nós da rede $\{k_i\} = \{k_1, ..., k_N\}$. Esta definição pode ser generalizada para um grau direcionado a sequências de graus de entrada $\{k_i^{in}\} = \{k_1^{in}, ..., k_N^{in}\}$, e para a sequência de graus de saída $\{k_i^{out}\} = \{k_1^{out}, ..., k_N^{out}\}$

4. **Grau médio:** dada a sequência de graus de um grafo não direcionado, o **grau médio** $\langle k \rangle$ é definido na forma:

$$\langle k \rangle = \frac{1}{N} \sum_i k_i = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ji} \tag{2.11}$$

5. **Distribuição de probabilidade de graus de um grafo:** o grau de um nó é uma propriedade local da rede, mas considerando a sequência de graus podemos determinar algumas propriedades globais da rede. A organização e estrutura geral de uma rede induzida por sua sequência de graus é caracterizada pela **distribuição de probabilidade de graus da rede *P(k)***. Para um grafo não direcionado, *P(k)* é definida como a fração de nós de grau k e representa a probabilidade de um nó escolhido aleatoriamente na rede ter grau k. Para um grafo direcionado, $P^{in}(k)$ e $P^{out}(k)$ são definidos respectivamente e analogamente como a fração de nós de grau de entrada k e a fração de nós de grau de saída k, que definem as distribuições de probabilidade de graus de entrada e saída. Ou seja, $P^{in}(k)$ e $P^{out}(k)$ representam as probabilidades de que a escolha aleatória de um nó na rede tenha respectivamente um grau de entrada k ou um grau de saída k.

6. **Caminho de um grafo:** um outro parâmetro importante é o"caminho em um grafo", o qual é definido como uma sequência de nós tal que dois nós consecutivos na sequência são conectados por uma aresta. Um caminho direcionado em uma rede direcionada é um caminho com arestas direcionadas de cada nó para o próximo na sequência. Cada caminho tem seu comprimento definido como o número de arestas que são passadas ao segui-lo, incluindo possíveis repetições em caminhos que também se cruzam. Os caminhos que partem de um nó e retornam a si mesmo no

final são chamados de **caminhos cíclicos** ou **ciclos**, enquanto aqueles que começam e terminam em nós diferentes são chamados de **caminhos acíclicos** ou **aciclos**. Caminhos acíclicos que visitam os nós do caminho apenas uma vez são chamados de **caminhos acíclicos auto-evitáveis**. Caminhos cíclicos que visitam os nós do caminho apenas uma vez, exceto o primeiro nó, também são chamados de **caminhos cíclicos auto-evitáveis**.

6. **Intermediação (*Betweenness*):** esta medida é baseada no número de menores caminhos que passam pelo vértice:

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \tag{2.12}$$

onde, σ_{st} é o número de menores caminhos entre s e t, e $\sigma_{st}(v_i)$ é o número de menores caminhos entre s e t passando por v_i.

Por tanto, a intermediação mede a centralidade de um vértice em termos de seu papel na conexão entre outros pares de vértices.

7. **Proximidade (*Closeness*):** mede quão próximo (central) é um vértice de todos os outros vértices da rede.

8. **Coeficiente de agrupamento (*Clustering coeficient*):** é uma medida do grau em que os nós de uma rede tendem a se agrupar. Dado um grafo não direcionado *G(N,L)*, é definido o coeficiente de agrupamento de um nó i do referido grafo e é denotado como C_i:

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{2.13}$$

9. ***Eigenvector Centrality*:** é uma medida baseada em auto-vetores.

10. ***Katz Centrality*:** é uma extensão da medida de "autovetores" para resolver o problema de caminhos com centralidade zero.

11. ***Page Rank*:** é uma das medidas mais importantes. Serve para resolver os problemas da medida *Katz*.

2.1.4 Redes regulares e aleatórias

As redes complexas estão localizadas entre redes regulares e redes totalmente aleatórias. Por tanto, vamos começar a sua descrição.

2.1.4.1 Redes regulares

Uma rede regular é definida como aquela em que todos os nós possuem o mesmo grau, ou seja, possuem o mesmo número de arestas incidentes e é denotada como uma rede ou grafo k-regular onde k é o grau de todos os nós. Na Figura 4 são apresentados exemplos de grafos k regulares.

Trivialmente, temos que um grafo completo K_n é um grafo regular *n-1*. As redes regulares aparecem em muitos sistemas naturais, como, por exemplo, na formação de estruturas cristalinas e obedecem a um princípio de regularidade. As redes regulares são caracterizadas por um coeficiente de agrupamento C muito alto (próximo de 1) e comprimento do caminho médio l também grande, especialmente para $k \ll N$. Na verdade $l \sim N/k$.

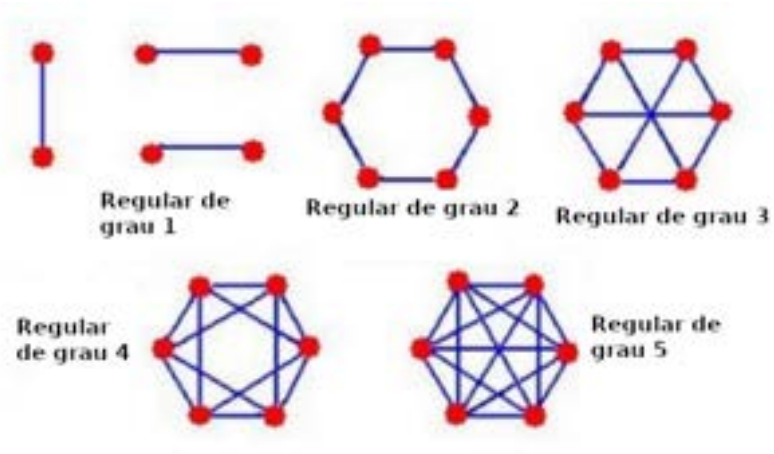

Figura 4 – Exemplos de grafos regulares (Universidad de Granada,).

2.1.4.2 Rede aleatória: rede de Erdös-Rényi (ER)

Uma rede aleatória é aquela em que o grau de cada nó é uma variável aleatória $k_i \in [0, N]$. O termo grafo aleatório refere-se à natureza aleatória da localização das arestas entre diferentes nós do grafo.
Erdös e Rényi (1960) propuseram um modelo para gerar grafos aleatórios com N nós e L arestas, que denotamos como rede aleatória de Erdös e Rényi (ER) ou grafo $G_{ER}(N, L)$. A maneira de construí-lo é a seguinte: começamos com N nós desconectados e começamos a conectar pares de nós selecionados aleatoriamente usando arestas, evitando a possibilidade de múltiplas arestas entre os mesmos nós, até que o número total de arestas atinja o valor L. Assim, o grafo resultante é apenas um dos muitos que podem ser gerados com N nós e L arestas. As redes aleatórias são caracterizadas por um coeficiente de agrupamento C relativamente baixo.

2.2 Estado da arte

Na previsão de "padrões de sucesso" que permitam visualizar as diferentes características que influenciam na aceitação de um livro por parte dos leitores, assim como prever os níveis de vendas e a posibilidade de futuras publicações, é possível a implementação de metódos de aprendizado de máquina e redes complexas para representar essas características e explorar os diferentes padrões presentes nas relações entre personagens.

Considernado algumas pesquisas previas, se encontrou que, Yucesoy *et al.* (2018) modelaram e analizaram a dinâmica da venda de livros, identificando uma série de padrões: **(1)** a maioria dos *bestsellers* atinge o pico de vendas em menos de dez semanas após o lançamento, **(2)** as vendas seguem um padrão universal de "pico precoce, decadência lenta", que pode ser descrito por um modelo estatístico preciso, **(3)** os autores mostraram que a fórmula prevista pelo modelo ajuda a prever as vendas futuras, mas para uma previsão precisa destas vendas é necessário pelo menos as primeiras 25 semanas de vendas após a publicação, um período dentro do qual a maioria dos livros atingiu seu pico de vendas e começou a perder ímpeto. Desta maneira, as previsões derivadas deste modelo estatístico são potencialmente úteis para o gerenciamento de estoque de longo prazo, mas não são eficazes para prever o potencial de vendas de um novo livro.

Recentemente foram propostas abordagens baseadas em redes complexas na análise de textos (Antiqueira *et al.*, 2007; Amancio *et al.*, 2011), partindo do fato que um texto pode ser representado pelas redes complexas com palavras conetadas por meio de procedimentos que dependem da sintaxe ou relações semânticas. As redes têm sido usadas para avaliar aspectos como a qualidade da escrita (Antiqueira *et al.*, 2007), reconhecimento de padrões na poesia e na prosa (Stevanak; Larue; Carr, 2010), assim como o estudo de propriedades gerais da linguagem escrita (Masucci; Rodgers, 2006). Mas, as redes de co-ocorrência, nas quais palavras adjacentes estão ligadas umas às outras, são provavelmente as mais populares para aplicações, devido à sua capacidade de capturar importantes aspectos sintáticos e semânticos de textos, com um procedimento de construção simples, se focando principalmente em escalas curtas, uma abordagem cada vez mais popular para escalas de texto mais longas (Herrera; Pury, 2008).

Amancio *et al.* (2011), realizaram uma quantificação de como as propriedades topológicas das redes de co-ocorrência de palavras e a intermitência na sua distribuição dependem da autoria, mostrando que os estilos de diferentes autores deixam impressões digitais em medidas estatísticas muito gerais de textos. Sendo que na avaliação da contribuição das medidas para o reconhecimento de autoria usaram três métodos de aprendizado de máquina. Os escores estatísticamente significativos obtidos na atribuição de autoria mostraram que a dependência de estilo dessas características pode ser usada na prática.

Por outro lado, Amancio (2015), estudou a eficácia da combinação de recursos textuais tradicionais e medidas topológicas de redes, no contexto de duas tarefas de processamento de linguagem natural: a atribuição de autoria e os problemas de identificação do gênero. Usando como classificadores para compor as técnicas de combinação: *K-Nearest-Neighbor* (kNN), *Support Vector Machines* (SVM), *Random Forest* (RFO) e *Multilayer Perceptron* (MLP). Em ambas as tarefas, a adição de características topológicas proporcionou uma melhora no desempenho da classificação.

Por todo o anteriormente planteado, ainda existe muito por melhorar e explorar na previsão de sucesso de um produto cultural tão importante como são os livros.

3 METODOLOGIA PARA PREVISÃO DE PADRÕES DE LIVROS

As características presentes no conteúdo dos livros que permitem determinar os padrões de sucesso dos mesmos, podem ser estudadas por meio de redes complexas seguindo uma quantificação estatística do papel das palavras nos textos (Amancio *et al.*, 2011). Embora os modelos atuais tenham sido úteis para revelar padrões por meio da análise de redes sintáticas e semânticas, apenas alguns trabalhos investigaram a importância da estrutura decorrente do relacionamento entre entidades relevantes, como personagens, locais e organizações. No caso desta pesquisa, cujo interesse é a determinação de padrões de succeso através das conexões entre personagens, foi usado o modelo proposto por Amancio (2016). Assim, foram realizadas as etapas de adquisição de dados, pré-processamento e processamento dos textos, criação das redes, caracterizacão das redes e finalmente a identificação de padrões. No presente capítulo esses processos são descritos de maneira geral, e no Capítulo 4 são abordados de maneira detalhada.

3.1 Dados

Para alcançar os objetivos propostos é necessário o conteúdo dos livros, sendo que o acesso a este conteúdo completo é disponibilizado apenas para livros antigos. Pelo que os dados usados nos experimentos pertencem a uma lista composta de romances de autores distintos, recuperados do repositório do *Project Gutenberg dataset*[1], uma biblioteca digital cujo acervo é composto por textos completos de aproximadamente 60 mil livros em domínio público em formato *eBook*.

A base de dados estudada é composta por 219 livros escritos em inglés, e foi dividida em duas cateorias: *sucesso* vs. *não sucesso*[2]. Para a categoria "*sucesso*" inicialmente as obras foram procuradas nas listas dos mais vendidos de *The New York Times Best Sellers*[3] (TNYTBS) e de *Publishers Weekly Bestseller Lists*[4] (PWBL); considerando apenas um livro por autor, a fim de evitar a identificação de autoria pelos algoritmos de *maching learning*. Finalmente a seleção de livros para compor a base de dados baseou-se na lista de *best sellers* da *Publishers Weekly*, devido ao número maior de obras disponíveis, no período comprendido entre 1895 e 1924. E na categoria de *não sucesso*, foram consideradas obras do repositorio, publicadas no mesmo período, mas que não foram inclusas na lista de *best sellers*. Na tabela 6 está ilustrada a lista completa de títulos das obras da categoria

[1] www.gutenberg.org.

[2] A base da dados foi construida previamente por Giovana Daniele da Silva, aluna da Universidade de São Paulo.

[3] https://www.nytimes.com/books/best-sellers/

[4] https://www.publishersweekly.com/pw/nielsen/index.html

"*sucesso*" inclusas na base de dados, assim como as informações de título, autoria e ano de publicação. E na tabela 7 estão ilustradas as informações das obras da categoria "*não sucesso*".

3.1.1 Pré-processamento do texto

Antes de extrair redes complexas e medidas de intermitência dos textos, algumas etapas de pré-processamento foram aplicadas:

1) Remoção de *stopwords*: processo no qual são eliminadas as palavras não relevantes para a aplicação, como artígos, preposições, pronomes e conetivos. Neste processo foi tratado o vetor de *embeddings* (mediante o modelo Doc2vec), a tokenização e a identificação das *stopwords*.

Algumas técnicas usadas no procesamento dos textos:

- Doc2Vec: modelo usado para criar uma representação vetorizada de um grupo de palavras tomadas coletivamente como uma única unidade.
- Análise de componentes básicos (PCA, *Principal Component Analysis*): é um procedimento estatístico que permite resumir o conteúdo das informações em grandes tabelas de dados por meio de um conjunto menor de "índices de resumo" que podem ser visualizados e analisados com mais facilidade.

2) Remoção de pontuação, letras maiúsculas e numeros;

3) Lematização (simplificação de termos): técnica para a redução de uma palavra a sua forma canônica e agrupação das diferentes formas da mesma palavra.

3.1.2 Processamento do texto

Após o pré-processamento, é possível realizar as fases que levam à construção das redes complexas dos livros, como a resolução de anáfora e o reconhecimento de entidades:

1) Resolução de anáfora:
Uma anáfora é um fenômeno linguístico que ocorre quando um elemento da fala se refere a algo já mencionado. Na resolução de anáfora, um classificador é treinado para decidir se um par de entidades nominais forma uma correferência. Então, cadeias de correferência completas são construídas agrupando essas decisões em pares.

2) Reconhecimento de entidade nomeada - NER:
Esta técnica identifica pessoas, locais e organizações relevantes em documentos. Após o reconhecimento de entidades, o texto passa por uma etapa de eliminação de inconsistencias e duplicações de entidades.

3.1.3 Criação das redes

A partir das informações obtidas do processameto dos textos, é possível construir a rede complexa para cada livro, e proseguir com a determinação de medidas de rede, e identificar os possíveis padrões presentes no comportamento das redes.

3.1.4 Medidas de rede

Depois das fases de pré-processamento e processamento, um cojunto de entidades $V = v_1, v_2, ...$ é obtido para cada livro. Para criar a rede de entidades relacionadas, referida como "rede de entidade nomeada (*Named Entity* - NE)", aquelas entidades que compartilham algum relacionamento semântico tendem a ser conectadas. A separação do texto em contextos é realizada dividindo-se todo o documento em subtextos mais curtos, compreendendo o mesmo número de tokens W. Como consequência, cada livro é representado pelo conjunto:

$$\Psi = \{S_1, S_2, ..., S_N\}, \tag{3.1}$$

com S_i sendo o subtexto *i-ésimo*. Para armazenar as informações relativas à co-ocorrência de entidades distintas no mesmo subconjunto, é criada a matriz $\mathbf{B}$. Se a entidade v_i aparacer no *j-ésimo* subconjunto, então $B_{ij} = 1$, caso contrário, $B_{ij} = 0$.

A frequência de entidades em subconjuntos, ou seja, o número de subtextos em que uma entidade aparece, é definido como $f_i = \sum B_{ij}$. Analogamente, a frequência de co-ocorrência de duas entidades v_i e v_j é definida como:

$$f_{ij} = \sum_k B_{ik} B_{jk}. \tag{3.2}$$

A ligação (aresta) entre duas entidades é estabelecida se elas co-ocorrem pelo menos em um conjunto $S_i \in \Psi$. O peso da aresta é calculado como:

$$w_{ij} = min\{P(v_i|v_j), P(v_j|v_i)\}, \tag{3.3}$$

onde $P(v_i|v_j) = f_{ij}/f_j$. Aquí, o peso w_{ij} é usado para identificar as arestas mais fortes, que por sua vez são armazanados na matriz de adjacência $\mathbf{A}$.

Para melhorar a caracterização das redes NE, também é considerada a significancia da co-ocorrência de entidades. Mais especificamente, duas entidades v_i e v_j são conectadas apenas se a quantidade f_{ij} é suficientemente maior do que o mesmo valor esperado em um modelo nulo, ou seja, em um texto aleatório.

Dadas duas entidades v_i e v_j, a significancia da co-ocorrência é estimada calculando a probabilidade de observar mais do que $k = f_{ij}$ co-ocorrências de v_i e v_j no modelo nulo. De forma equivalente, o *valor p* associado à quantidade $k = f_{ij}$ é $p = \sum_{k \geq r} p(k)$, onde $p(k)$ é a probabilidade de k co-ocorrências de v_i e v_j no texto aleatório. Para calcular $p(k)$, é seguida a abordagem:

Se $n_1 = f_i$ e $n_2 = f_j$, $p(k)$ pode ser calculado como:

$$p(k) = \frac{(N; k, n_1 - k, n_2 - k)}{(N; n_1)(N; n_2)}, \tag{3.4}$$

onde $(x; y_1, ..., y_n)$ é a notação simplificada:

$$(x; y_1, ..., y_n) \equiv \frac{x!}{y_1!...y_n!} \frac{1}{(x - y_1 - ...y_n)!}. \tag{3.5}$$

A Equação 3.5 pode ser reescrita de uma maneira mais conveniente, se a notação $\{a\}_b$ definida como:

$$\{a\}_b \equiv \prod_{i=0}^{b-1} (a - i), \tag{3.6}$$

é adotada para $a \geq b$. Neste caso, a probabilidade $p(k)$ pode ser escrita como:

$$\begin{aligned} p(k) &= \frac{\{n_1\}_k \{n_2\}_k \{N - n_1\}_{n_2 - k}}{\{N\}_{n_2} \{k\}_k} = \frac{\{n_1\}_k \{n_2\}_k \{N - n_1\}_{n_2 - k}}{\{N\}_{n_2 - k} \{N - n_2 + k\}_k \{k\}_k} \\ &= \prod_{j=0}^{n_2 - k - 1} \left[\frac{N - j - n_1}{N - j} \right] \times \prod_{j=0}^{k-1} \frac{(n_1 - j)(n_2 - j)}{(N - n_2 + k - j)(k - j)}. \end{aligned} \tag{3.7}$$

Por tanto o valor p associado ao número de co-ocorrências observadas é:

$$p(k) = \sum_{k \geq r} \prod_{j=0}^{n_2 - k - 1} \left(1 - \frac{n_1}{N - j} \right) \times \prod_{j=0}^{k-1} \frac{(n_1 - j)(n_2 - j)}{(N - n_2 + k - j)(k - j)}. \tag{3.8}$$

Esse valor $p(k)$ pode ser usado para estabelecer arestas entre entidades cuja frequência de co-ocorrência é significativa.

3.1.4.1 Caracterização da rede

As redes são caracterizadas pelas propriedades de centralidade, que se refere à posição dos nós nas redes, e de centralização, entendida como toda a estrutura de uma rede. Por sua vez, essas noções baseiam-se nos conceitos de grau, intermediação e proximidade; além disso, toda rede possui uma densidade.

3.1.4.2 Identificação de padrões

Após a cálculo das medidas de centralidade, é feita uma identificação dos padrões presentes na rede. No nosso caso, o foco está no relacionamento entre as pesonagens, pelo que é determinado se esse relacionamento realmente tem um peso no sucesso de vendas e aceitação dos livros.

4 EXPERIMENTOS E ANÁLISE DE RESULTADOS

As redes foram construidas a partir da identificação das personagens nos livros. Mas, para o propósito da categorização dos dados, para a qual foi usada a "**classificação supervisionada**[1]", as duas amostras que constituem a base de dados: **sucesso** (no caso dos *best seller*) e **não sucesso**, foram unificadas. Assim, na fase de processamento e treinamento dos modelos, obtivemos uma amostra de 219 livros (correspondente a base de dados inicial - Seção 3.1). A partir desta nova amostra, e após o pré-processamento dos dados (remoção de *stopwords*, pontuação, letras maiúsculas, numeros, etc.), se realizaram os seguintes procesos:

4.1 Reconhecimento de entidades

Esta etapa consistiu em identificar as personagens presentes nos livros. No caso da amostra completa (219 livros), a identificação foi feita por parágrafos. Na Tabela 1 são apresentados os resultados deste processo para 10 livros da amostra, onde é possivel ver o número de parágrafos identificados para cada livro (coluna 3) e o número de parágafos contendo personagens (coluna 4).

Tabela 1 – Resultados do reconhecimento de entidades para 10 livros da amostra.

Livro	Autor	Número parágrafos	Número parágrafos com personagens
Beside The Bonnie Brier Bush	Ian Maclaren	1358	1158
Trilby	George Du Maurier	2407	1987
The Adventures Of Captain Horn	Frank Richard Stockton	1945	1608
The Manxman	Hall Caine	5701	5227
The Princess Aline	Richard Harding Davis	431	357
The Master	Israel Zangwill	3962	3511
The Prisoner Of Zenda	Anthony Hope	1762	1538
Degeneration	Max Nordau	1835	1346
My Lady Nobody	Maarten Maartens	4095	3714
Tom Grogan	Francis Hopkinson Smith	803	695

4.2 Criação da rede de co-ocorrência

As redes de co-ocorrência são geralmente usadas para fornecer uma exibição gráfica de possíveis relacionamentos entre pessoas, organizações, conceitos ou outras entidades representadas em material escrito. Por tanto, e como foi comentado na seção 2.2, as redes de co-ocorrência permitiram identificar as relações de uma palavra (personagem) chave com outras palavras chave ou com os grupos que formam essas palavras. Aquí foi feita uma

[1] Na classificação supervisionada, uma relação binária mapeando a entrada para uma saída é gerada pelo algoritmo.

interação por parâgrafos, a fim de gerar a **matriz de co-ocorrência de personagens** e o **conjunto de arestas entre as personagens**.

4.2.1 Visualização de rede

Foram gerados os gráficos de redes aleatórias a partir dos dados na matriz de co-ocorrência[2]. Nos Apêndices B (Figuras 8, 9, 10, 11, 12, 13, 14, 15, 16 e 17) e C (Figuras 18, 19, 20, 21, 22, 23, 24, 25, 26 e 27), é possível observar esas redes, tanto para a classe "*sucesso*" quanto para a classe "*não sucesso*", sendo possível comprender visualmente as relações entre as personagens. Porém, o comportamento das redes para ambas as classes são similares, tornando pouco possível determinar um padrão específico para as classes que levem a inferir uma diferença entre o relaçonamento das personagens que possa influenciar o sucesso dos livros.

4.2.1.1 Medidas de centralidade

Na determinação de medidas de rede, primeiro foram feitos vários testes, a fim de análisar e escolher aquelas que foram mais convenientes para as amostras, pois algumas delas não tinham um impacto importante na análise. Além disso, as medidas foram determinadas com peso e sem peso, mas os resultados obtidos foram quase iguais, assim que optamos pelas medidas **sem peso**. Finalmente, para a análises dos dados foram elegidas as medidas de centralidade registradas na Tabela 2.

	Nome da medida
Medida 1	Centralidade de grau
Medida 2	*Clustering coefficient*
Medida 3	*Closeness centrality*
Medida 4	*Betweenness centrality*
Medida 5	*Eigenvetor centrality*
Medida 6	*PageRank*

Tabela 2 – Medidas usadas na análise dos dados.

As medidas foram determinadas para todos os livros da amostra total. Mas, como os resultados são muito extensos, aquí são apresentados os resultados só para 15 livros da classe **sucesso** (Tabela 3) e 15 livros da classe **não sucesso** (Tabela 4). Por outro lado, para uma interpretação certa dos resultados é necessário comprender os modelos escolhidos:

[2] Implementado com networkX

- **KNN (K vizinhos mais próximos):** é um classificador de aprendizagem supervisionado não paramétrico, que usa proximidade para fazer classificações ou previsões sobre o agrupamento de um ponto de dados individual.
- ***Random Forest*:** é uma técnica de aprendizagem supervisionada que gera múltiplas árvores de decisão em um conjunto de dados de treinamento. Os resultados obtidos são combinados para obter um único modelo mais robusto comparado aos resultados de cada árvore separadamente,
- **SVM (*Support Vector Method*):** é um novo método geral de estimativa de função que não depende explicitamente da dimensionalidade do espaço de entrada. É aplicado para problemas de reconhecimento de padrões, estimativa de regressão e estimativa de densidade, bem como para problemas de resolução de equações de operadores lineares (VAPNIK, 1997).

A correlação entre as medidas foram feitas a través de mapas de calor para as métricas: KNN (Figura 5), *Random Forest* (Figura 6) e SVM (Figura 7), e suas respectivas matrizes de confusão. Lembrando que uma **matriz de confusão**, é uma ferramenta importante para avaliar o desempenho de modelos de classificação. Ela mostra a contagem de verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN).
As três métricas são análisadas a través dos parâmetros:

- **Precisão:** fornece a precisão da medida,
- ***Recall*:** indica a proporção de exemplos positivos,
- **F1-*Score*:** sendo o mais importante nos resultados aquí. Em um modelo de classificação binária, um valor grande de F1-Score (~ 1), indica excelente precisão e recall, enquanto um valor baixo indica desempenho insatisfatório do modelo. Assim, um F1-*Score* mais alto sugere melhor desempenho do modelo.

As cores mais obscuras nos mapas de calor, indicam maior correlação. Adicionalmente, os valores numéricos foram registrados na Tabela 5.

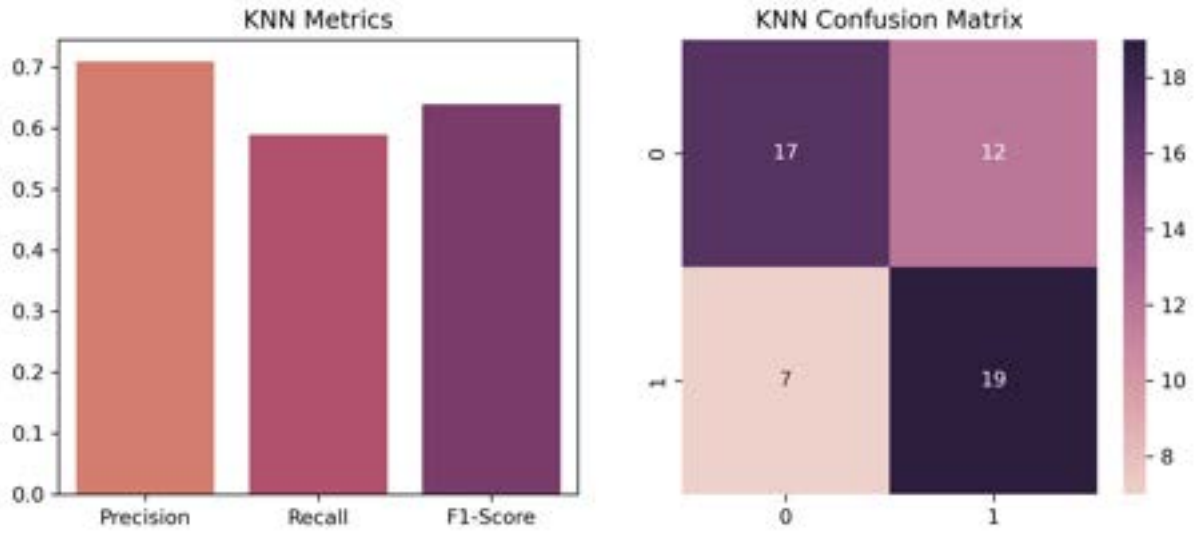

Figura 5 – Mapa de calor e matriz de confusão para a métrica KNN.

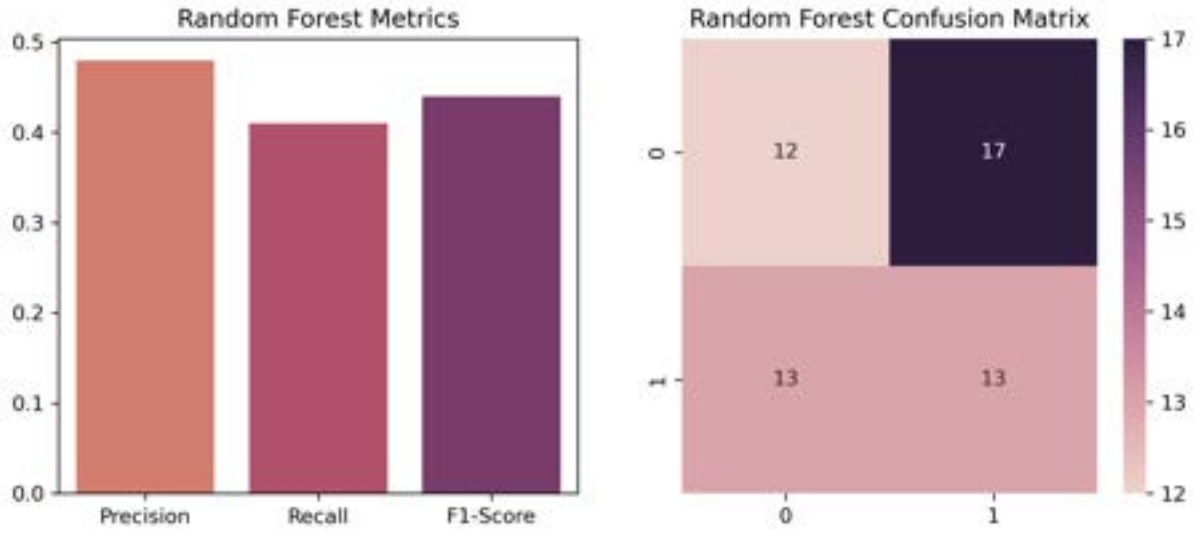

Figura 6 – Mapa de calor e matriz de confusão para a métrica *Random Forest*.

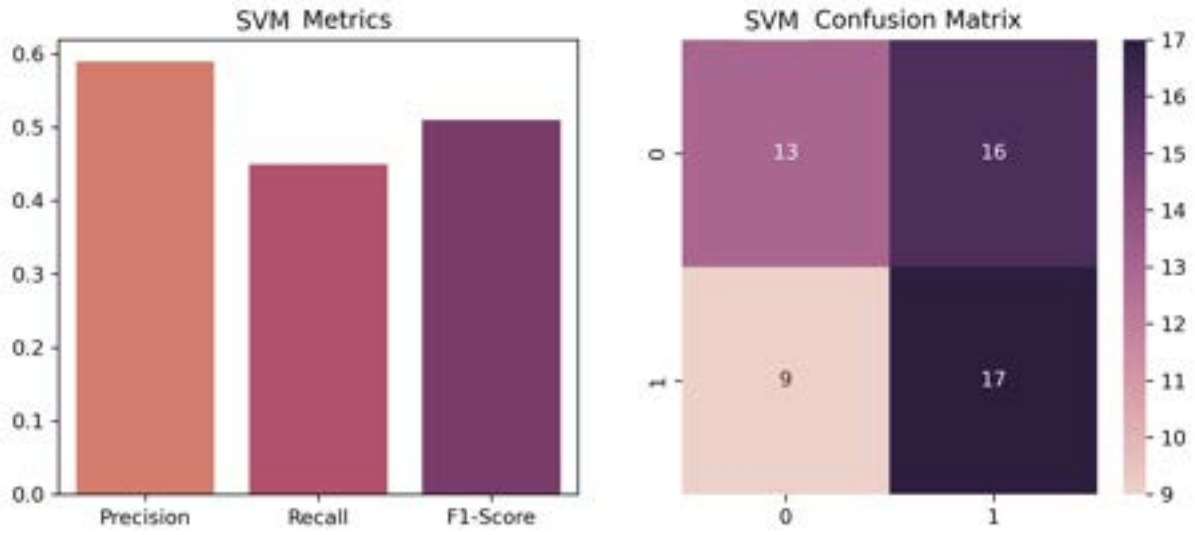

Figura 7 – Mapa de calor e matriz de confusão para o SVM.

Tabela 3 – Medidas determinadas para uma subamostra de 15 livros da **classe 1: sucesso**.

Nome do livro	Autor	Ano	Classe	Medida 1	Medida 2	Medida 3	Medida 4	Medida 5	Medida 6
Beside the Bonnie Brier Bush	Ian Maclaren	1895	1	0.0169	0.5135	0.2145	0.0087	0.0327	0.0033
Trilby	George Du Maurier	1895	1	0.0079	0.5081	0.1555	0.0040	0.0212	0.0020
The Adventures of Captain Horn	Frank Richard Stockton	1895	1	0.0425	0.5156	0.2746	0.0241	0.0630	0.0095
The Manxman	Hall Caine	1895	1	0.0088	0.4963	0.2449	0.0042	0.0220	0.0017
The Princess Aline	Richard Harding Davis	1895	1	0.0706	0.5021	0.3586	0.0441	0.1009	0.0185
The Master	Israel Zangwill	1895	1	0.0123	0.5471	0.2227	0.0043	0.0225	0.0018
The Prisoner of Zenda	Anthony Hope	1895	1	0.0348	0.3903	0.1941	0.0147	0.0492	0.0076
Degeneration	Max Nordau	1895	1	0.0060	0.6497	0.0844	0.0018	0.0097	0.0010
My Lady Nobody	Maarten Maartens	1895	1	0.0075	0.3038	0.1645	0.0047	0.0244	0.0024
Tom Grogan	Francis Hopkinson Smith	1896	1	0.0556	0.6394	0.3641	0.0215	0.0597	0.0074
A Lady of Quality	Frances Hodgson Burnett	1896	1	0.0284	0.4095	0.1975	0.0176	0.0550	0.0085
The Seats of the Mighty	Gilbert Parker	1896	1	0.0104	0.3912	0.0946	0.0056	0.0299	0.0038
A Singular Life	Elizabeth Stuart Phelps Ward	1896	1	0.0153	0.4399	0.1578	0.0072	0.0307	0.0038
The Damnation of Theron Ware	Harold Frederic	1896	1	0.0176	0.5501	0.1551	0.0073	0.0361	0.0043
A House-Boat on the Styx	John Kendrick Bangs	1896	1	0.0232	0.4623	0.2243	0.0169	0.0470	0.0060

Tabela 4 – Medidas determinadas para uma subamostra de 15 livros da **classe 0: não sucesso**.

Nome do livro	Autor	Ano	Classe	Medida 1	Medida 2	Medida 3	Medida 4	Medida 5	Medida 6
In the Land Of Cave And Cliff Dwellers	Frederick Schwatka	1895	0	0.0289	0.4007	0.0596	0.0056	0.0443	0.0135
Jude the Obscure	Thomas Hardy	1895	0	0.0114	0.3819	0.1173	0.0050	0.0320	0.0042
The Golden Age	Kenneth Grahame	1895	0	0.0293	0.6009	0.2385	0.0200	0.0563	0.0083
The Lost Stradivarius	John Meade Falkner	1895	0	0.0753	0.6381	0.3528	0.0297	0.0808	0.0127
The British Barbarians	Grant Allen	1895	0	0.0632	0.5064	0.2791	0.0310	0.0812	0.0147
The Sorrows of Satan	Marie Corelli	1895	0	0.0128	0.4143	0.1378	0.0051	0.0274	0.0031
Rose of Dutcher's Coolly	Hamlin Garland	1895	0	0.0162	0.4007	0.1162	0.0074	0.0403	0.0057
Yekl: A Tale of the New York Ghetto	Abraham Cahan	1896	0	0.0482	0.4955	0.2666	0.0234	0.0679	0.0111
Madelon	Mary E. Wilkins Freeman	1896	0	0.0432	0.5565	0.3061	0.0184	0.0535	0.0074
The Country of the Pointed Firs	Sarah Orne Jewett	1896	0	0.0350	0.5025	0.2325	0.0191	0.0622	0.0095
The Well at the World's End	William Morris	1896	0	0.0166	0.5725	0.2924	0.0065	0.0303	0.0027
A Child of the Jago	Arthur Morrison	1896	0	0.0430	0.5630	0.3083	0.0151	0.0495	0.0057
Tom Sawyer, Detective	Mark Twain	1896	0	0.0736	0.4618	0.2487	0.0316	0.0802	0.0141
In His Steps	Charles M. Sheldon	1896	0	0.0363	0.5265	0.2819	0.0168	0.0471	0.0063
History of the Warfare of Science...	Andrew Dickson White	1896	0	0.0035	0.5789	0.1183	0.0017	0.0102	0.0007

Tabela 5 – Resultados da avaliação dos modelos para ambas as clases: 1 (sucesso) e 0 (não sucesso).

Modelo	Acurácia	Precisão		Recall		F1-Score		Support	
		1	0	1	0	1	0	1	0
KNN	0.65	0.61	0.71	0.73	0.59	0.67	0.64	26	29
Random Forest	0.45	0.43	0.48	0.50	0.41	0.46	0.44	26	29
SVC	0.55	0.52	0.59	0.65	0.45	0.58	0.51	26	29

4.2.1.2 Vectorização das medidas de centralidade de rede:

Após a determinação das medidas de rede, foi gerado um vetor com as médias para cada livro, e esses vetores foram agrupados por classe: **classe 0** para os livros da categoría "não sucesso" e **classe 1** para os livros da categoría "sucesso". Esses dados são usados na etapa final de classificação.

4.2.2 Classificação

Nesta fase são avaliados os modelos e obtidos os resultados finais. A partir dos dados obtidos na fase anterior, foram determinados dois conjuntos de dados: **treinamento** (75%) e **teste** (25%).

4.2.2.1 Avaliação dos modelos

Após avaliar os modelos, e a partir dos valores obtidos dos parâmetros: Acurácia, Precisão, *Recall*, F1-*Score* e o *Support*, apresentados na Tabela 5, observamos que a acurácia é $\geqslant 45\%$; a presição para ambas as clases (1: sucesso, 0: não sucesso) fornece uma diferença, sendo maior para a classe 0; já no *Recall* aconteceu o contrario, identificando uma maior proporção de exemplos positivos para a classe 1. E como foi mencionado anteriormente, o F1-*Score* representa a métrica mais importante, fornecendo um maior desempenho do modelo KNN para ambas as classes, e identificando mais os livros que são sucesso (classe 1).

5 CONCLUSÃO

Após realizar todo o proceso de pre-processamento e processamento dos dados, observamos que quando se fala em sucesso de um livro qualquer, é necessário analisar todos os fatores que influenciam esse sucesso, pois a partir dos resultados obtidos (ver Tabela 5), podemos deduzir que o relacionamento entre as personagens dos livros não é suficiente para prever que o livro em questão obterá o sucesso desejado. Por tanto, ainda é preciso estudar outros parâmetros que posibilitem uma análise mais robusta, já que um só não consigue medir tal sucesso.

5.1 Trabalhos futuros

Os resultados obtidos neste projeto permitem a continuação da pesquisa, quanto a exploração de outros parâmetros que podem influenciar no sucesso ou não sucesso de livros. Assim, ainda há muito por explorar nesta área, desenvolvendo o potencial das redes complexas em um conceito que envolve un produto cultural tão importante como a leitura de livros, e ajudando a industria da venda de livros na previsão de bons investimentos.

REFERÊNCIAS

Amancio, D. R. A complex network approach to stylometry. **PLoS ONE**, IOP Publishing, Aug 2015. Available at: https://doi.org/10.1371/journal.pone.0136076.

Amancio, D. R. Network analysis of named entity co-occurrences in written texts. IOP Publishing, v. 114, n. 5, p. 58005, jun 2016. Available at: https://doi.org/10.1209/0295-5075/114/58005.

Amancio, D. R. *et al.* Comparing intermittency and network measurements of words and their dependence on authorship. IOP Publishing, v. 13, n. 12, p. 123024, dec 2011. Available at: https://doi.org/10.1088/1367-2630/13/12/123024.

Antiqueira, L. *et al.* Strong correlations between text quality and complex networks features. **Physica A: Statistical Mechanics and its Applications**, v. 373, p. 811–820, 2007. ISSN 0378-4371. Available at: https://www.sciencedirect.com/science/article/pii/S0378437106006881.

Beck, J. The sales effect of word of mouth: a model for creative goods and estimates for novels. **J Cult Econ**, v. 31, p. 5–23, 2007. Available at: https://doi.org/10.1007/s10824-006-9029-0.

Clement, M.; Proppe, D.; Rott, A. Do critics make bestsellers? opinion leaders and the success of books. **Journal of Media Economics**, Routledge, v. 20, n. 2, p. 77–105, 2007. Available at: https://doi.org/10.1080/08997760701193720.

Herrera, J. P.; Pury, P. A. Statistical keyword detection in literary corpora. **The European Physical Journal B**, v. 63, 2008.

Masucci, A. P.; Rodgers, G. J. Network properties of written human language. **Phys. Rev. E**, American Physical Society, v. 74, p. 026102, Aug 2006. Available at: https://link.aps.org/doi/10.1103/PhysRevE.74.026102.

Nakamura, L. "words with friends": Socially networked reading on goodreads. **PMLA/Publications of the Modern Language Association of America**, Cambridge University Press, v. 128, n. 1, p. 238–243, 2013.

Schmidt-Stölting, C.; Blömeke, E.; Clement, M. Success drivers of fiction books: An empirical analysis of hardcover and paperback editions in germany. **Journal of Media Economics**, Routledge, v. 24, n. 1, p. 24–47, 2011. Available at: https://doi.org/10.1080/08997764.2011.549428.

Shehu, E. *et al.* The influence of book advertising on sales in the german fiction book market. **J Cult Econ**, Routledge, v. 38, p. 109–130, 2014. Available at: https://doi.org/10.1007/s10824-013-9203-0.

Stevanak, J. T.; Larue, D. M.; Carr, L. D. Distinguishing fact from fiction: Pattern recognition in texts using complex networks. 2010.

Universidad de Granada. **www.ugr.es, 01 de Marzo de 2023**.

VAPNIK, V. N. The support vector method. *In*: GERSTNER, W. *et al.* (ed.). **Artificial Neural Networks — ICANN'97**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. p. 261–271. ISBN 978-3-540-69620-9.

Wang, X. *et al.* Success in books: predicting book sales before publication. **EPJ Data Science**, v. 8, p. 1–20, out. 2019.

Yucesoy, B. *et al.* Success in books: a big data approach to bestsellers. **EPJ Data Science**, v. 7, Apr 2018.

APÊNDICES

APÊNDICE A – TABELAS

Tabela 6 – Lista das 110 obras da categoria *sucesso* usadas na construção da base de dados.

Título	**Autor(a)**	**Ano de publicação**
Beside the Bonnie Brier Bush	Ian Maclaren	1895
Trilby	George Du Maurier	1895
The Adventures of Captain Horn	Frank Richard Stockton	1895
The Manxman	Hall Caine	1895
The Princess Aline	Richard Harding Davis	1895
The Master	Israel Zangwill	1895
The Prisoner of Zenda	Anthony Hope	1895
Degeneration	Max Nordau	1895
My Lady Nobody	Maarten Maartens	1895
Tom Grogan	Francis Hopkinson Smith	1896
A Lady of Quality	Frances Hodgson Burnett	1896
The Seats of the Mighty	Gilbert Parker	1896
A Singular Life	Elizabeth Stuart Phelps Ward	1896
The Damnation of Theron Ware	Harold Frederic	1896
A House-Boat on the Styx	John Kendrick Bangs	1896
The Red Badge of Courage	Stephen Crane	1896
Sentimental Tommy	J. M. Barrie	1896
Quo Vadis	Henryk Sienkiewicz	1897
The Choir Invisible	James Lane Allen	1897
On the Face of the Waters	Flora Annie Steel	1897
The Honorable Peter Stirling	Paul Leicester Ford	1897
Hugh Wynne	Silas Weir Mitchell	1898
Penelope's Progress	Kate Douglas Wiggin	1898
Helbeck of Bannisdale	Mary Augusta Ward	1898
The Pride of Jennico	Egerton Castle	1898
Shrewsbury	Stanley J. Weyman	1898

Continua na página seguinte.

Título	Autor(a)	Ano de publicação
David Harum	Edward Noyes Westcott	1899
When Knighthood Was in Flower	Charles Major	1899
Richard Carvel	Winston Churchill	1899
Red Rock	Thomas Nelson Page	1899
Aylwin	Theodore Watts-Dunton	1899
Mr. Dooley in Peace and War	Finley Peter Dunne	1899
To Have and to Hold	Mary Johnston	1900
Red Pottage	Mary Cholmondeley	1900
Unleavened Bread	Robert Grant	1900
Eben Holden	Irving Bacheller	1900
The Redemption of David Corson	Charles Frederic Goss	1900
Alice of Old Vincennes	Maurice Thompson	1900
The Helmet of Navarre	Bertha Runkle	1901
The Puppet Crown	Harold MacGrath	1901
The Life and Death of Richard Yea-and-Nay	Maurice Hewlett	1901
Graustark	George Barr McCutcheon	1901
The Virginian	Owen Wister	1902
Mrs. Wiggs of the Cabbage Patch	Alice Hegan Rice	1902
The Mississippi Bubble	Emerson Hough	1902
The Hound of the Baskervilles	Arthur Conan Doyle	1902
The Two Vanrevels	Booth Tarkington	1902
The Blue Flower	Henry van Dyke	1902
Sir Richard Calmady	Lucas Malet	1902
The Pit	Frank Norris	1903
The One Woman	Thomas Dixon Jr.	1903
The Little Shepherd of Kingdom Come	John Fox Jr.	1903
The Deliverance	Ellen Glasgow	1904
The Masquerader	Katherine Cecil Thurston	1904
In the Bishop's Carriage	Miriam Michelson	1904
My Friend Prospero	Henry Harland	1904

Continua na página seguinte.

Título	**Autor(a)**	**Ano de publicação**
The Silent Places	Stewart Edward White	1904
The Garden of Allah	Robert Hichens	1905
The House of Mirth	Edith Wharton	1905
The Princess Passes	C. N. Williamson	1905
The Fighting Chance	Robert W. Chambers	1906
The House of a Thousand Candles	Meredith Nicholson	1906
The Jungle	Upton Sinclair	1906
The Awakening of Helena Richie	Margaret Deland	1906
The Spoilers	Rex Beach	1906
The Brass Bowl	Louis Joseph Vance	1907
Satan Sanderson	Hallie Erminie Rives	1907
The Doctor	Ralph Connor	1907
The Inner Shrine	Basil King	1909
Katrine	Elinor Macartney Lane	1909
The Man in Lower Ten	Mary Roberts Rinehart	1909
Septimus	William J. Locke	1909
The Rosary	Florence L. Barclay	1910
Molly Make-Believe	Eleanor Hallowell Abbott	1910
The Broad Highway	Jeffery Farnol	1911
The Prodigal Judge	Vaughan Kester	1911
The Winning of Barbara Worth	Harold Bell Wright	1911
Queed	Henry Sydnor Harrison	1911
The Harvester	Gene Stratton Porter	1911
The Melting of Molly	Maria Thompson Daviess	1912
Tante	Anne Douglas Sedgwick	1912
Fran	J. Breckenridge Ellis	1912
Pollyanna	Eleanor H. Porter	1913
The Salamander	Owen Johnson	1914
Diane of the Green Van	Leona Dalrymple	1914
The Devil's Garden	W. B. Maxwell	1914
The Harbor	Ernest Poole	1915
The Lone Star Ranger	Zane Grey	1915
Mr. Britling Sees It Through	H. G. Wells	1916

Continua na página seguinte.

Título	Autor(a)	Ano de publicação
The Real Adventure	Henry Kitchell Webster	1916
Bars of Iron	Ethel M. Dell	1916
Nan of Music Mountain	Frank H. Spearman	1916
The Heart of Rachael	Kathleen Norris	1916
The Tree of Heaven	May Sinclair	1918
The Pawns Count	E. Phillips Oppenheim	1918
Sonia	Stephen McKenna	1918
The Arrow of Gold	Joseph Conrad	1919
The Tin Soldier	Temple Bailey	1919
Kindred of the Dust	Peter B. Kyne	1920
The River's End	James Oliver Curwood	1920
The Portygee	Joseph C. Lincoln	1920
Main Street	Sinclair Lewis	1921
The Brimming Cup	Dorothy Canfield	1921
The Sheik	Edith M. Hull	1921
The Sisters-in-Law	Gertrude Atherton	1921
The Kingdom Round the Corner	Coningsby Dawson	1921
If Winter Comes	A. S. M. Hutchinson	1922
Simon Called Peter	Robert Keable	1922
Maria Chapdelaine	Louis Hémon	1922
The Sea Hawk	Rafael Sabatini	1922

Fonte: Autoria própria.

Tabela 7 – Lista das 109 obras da categoria *não sucesso* usadas na construção da base de dados.

Título	Autor(a)	Ano de publicação
In the Land of Cave and Cliff Dwellers	Frederick Schwatka	1895
Jude the Obscure	Thomas Hardy	1895
The Golden Age	Kenneth Grahame	1895
The Lost Stradivarius	John Meade Falkner	1895
The British Barbarians	Grant Allen	1895

Continua na página seguinte.

Título	**Autor(a)**	**Ano de publicação**
The Sorrows of Satan	Marie Corelli	1895
Rose of Dutcher's Coolly	Hamlin Garland	1895
Yekl: A Tale of the New York Ghetto	Abraham Cahan	1896
Madelon	Mary E. Wilkins Freeman	1896
The Country of the Pointed Firs	Sarah Orne Jewett	1896
The Well at the World's End	William Morris	1896
A Child of the Jago	Arthur Morrison	1896
Tom Sawyer, Detective	Mark Twain	1896
In His Steps	Charles M. Sheldon	1896
History of the Warfare of Science with Theology in Christendom	Andrew Dickson White	1896
The Whirlpool	George Gissing	1897
The Beth Book	Sarah Grand	1897
The Spoils of Poynton	Henry James	1897
The Beetle	Richard Marsh	1897
Added Upon	Nephi Anderson	1898
A Digit of the Moon	F. W. Bain	1898
A Man from the North	Arnold Bennett	1898
The Uncalled	Paul Laurence Dunbar	1898
The Second Thoughts of an Idle Fellow	Jerome K. Jerome	1898
Evelyn Innes	George Moore	1898
The Golden Canyon	G. A. Henty	1899
Some Experiences of an Irish R.M.	Somerville and Ross	1899
Wood and Garden	Gertrude Jekyll	1899
The Theory of the Leisure Class	Thorstein Veblen	1899
A Book of the West	Sabine Baring-Gould	1899
The Wonderful Wizard of Oz	L. Frank Baum	1900
The Infidel	Mary Elizabeth Braddon	1900
The Wallet of Kai Lung	Ernest Bramah	1900

Continua na página seguinte.

Título	**Autor(a)**	**Ano de publicação**
Sister Carrie	Theodore Dreiser	1900
The Brass Bottle	Thomas Anstey Guthrie	1900
Erewhon Revisited	Samuel Butler	1901
The House with the Green Shutters	George Douglas Brown	1901
My Brilliant Career	Miles Franklin	1901
The Purple Cloud	M. P. Shiel	1901
Up From Slavery	Booker T. Washington	1901
The Sheep-Stealers	Violet Jacob	1902
A Daughter of the Snows	Jack London	1902
The Four Feathers	A. E. W. Mason	1902
The Flight of Pony Baker	W. D. Howells	1902
The Riddle Of The Sands	Erskine Childers	1903
Long Will	Florence Converse	1903
The Jewel of Seven Stars	Bram Stoker	1903
Thirty Years in Australia	Ada Cambridge	1903
The Napoleon of Notting Hill	Gilbert K. Chesterton	1904
The Island Pharisees	John Galsworthy	1904
Cabbages and Kings	O. Henry	1904
The Common Lot	Robert Herrick	1904
Green Mansions	W. H. Hudson	1904
A Diary from Dixie	Mary Boykin Miller Chesnut	1905
Where Angels Fear to Tread	E. M. Forster	1905
The Scarlet Pimpernel	Baroness Orczy	1905
The Education of Henry Adams	Henry Adams	1906
The Fifth Queen	Ford Madox Ford	1906
Romance Island	Zona Gale	1906
Love Among the Chickens	P. G. Wodehouse	1906
The Fortunes of Philippa	Angela Brazil	1906
Pip	Ian Hay	1907
The Mystery of “The Yellow Room”	Gaston Leroux	1907
The Hill of Dreams	Arthur Machen	1907
Father and Son	Edmund Gosse	1907
Jimbo	Algernon Blackwood	1909

Continua na página seguinte.

Título	Autor(a)	Ano de publicação
The Promise Of American Life	Herbert David Croly	1909
Elusive Isabel	Jacques Futrelle	1909
Multitude and Solitude	John Masefield	1909
Prester John	John Buchan	1910
The Return	Walter de la Mare	1910
Zuleika Dobson	Max Beerbohm	1911
The Hampdenshire Wonder	John Davys Beresford	1911
The Quest of the Silver Fleece	W. E. B. Du Bois	1911
Dawn O'Hara	Edna Ferber	1911
The Eye of Osiris	R. Austin Freeman	1911
The Promised Land	Mary Antin	1912
Mrs. Ames	E. F. Benson	1912
Mates at Billabong	Mary Grant Bruce	1912
Trent's Last Case	Edmund Clerihew Bentley	1913
The King of Alsander	James Elroy Flecker	1914
Bird of Paradise	Ada Leverson	1914
I Pose	Stella Benson	1915
Song of the Lark	Willa Cather	1915
Windy McPherson's Son	Sherwood Anderson	1916
Mendel	Gilbert Cannan	1916
Metamorphosis	Franz Kafka	1916
Backwater	Dorothy M. Richardson	1916
Seven Miles to Arden	Ruth Sawyer	1916
The Beasts of Tarzan	Edgar Rice Burroughs	1916
Regiment of Women	Clemence Dane	1917
Meccania	Owen Gregory	1918
Patricia Brent, Spinster	Herbert Jenkins	1918
Tarr	Percy Wyndham Lewis	1918
The Valley of Squinting Windows	Brinsley MacNamara	1918
The Young Visiters	Daisy Ashford	1919
Jurgen	James Branch Cabell	1919
The Mysterious Affair at Styles	Agatha Christie	1920
Alf's Button	W. A. Darlington	1920
Three Soldiers	John Dos Passos	1920

Continua na página seguinte.

Título	**Autor(a)**	**Ano de publicação**
Mystery Ranch	Arthur Chapman	1921
The Beautiful and Damned	F. Scott Fitzgerald	1921
She and Allan	H. Rider Haggard	1921
The Black Moth	Georgette Heyer	1921
Crome Yellow	Aldous Huxley	1921
Ulysses	James Joyce	1922
The Love Story of Aliette Brunton	Gilbert Frankau	1922
Lady Into Fox	David Garnett	1922
Whose Body?	Dorothy L. Sayers	1923

Fonte: Autoria própria.

APÊNDICE B – GRÁFICOS: LIVROS CLASSIFICAÇÃO "SUCESSO"

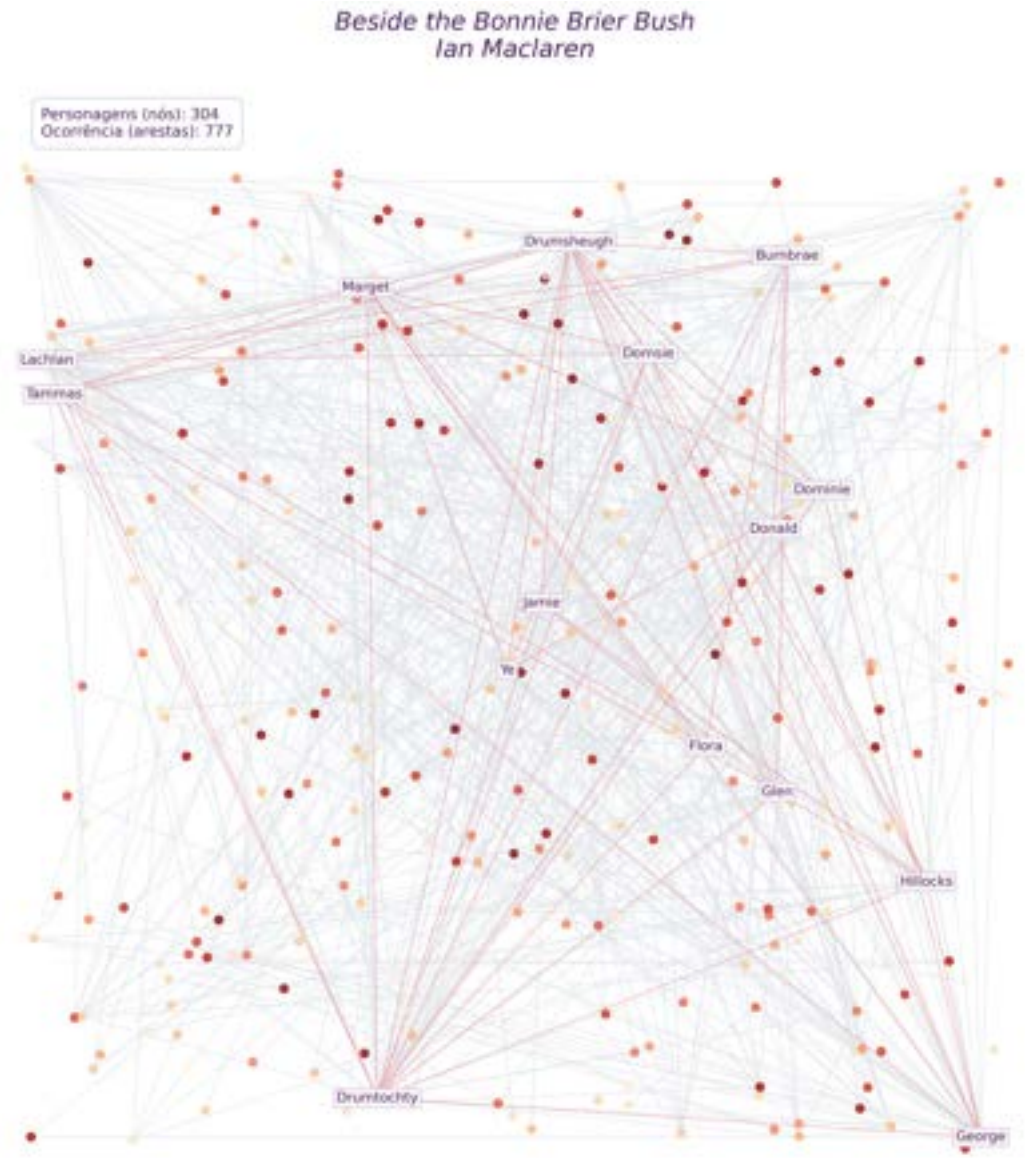

Figura 8 – Rede aleatória construida para o livro: *Beside the Bonnie Brier Bush*, pertencente à classificação "sucesso".

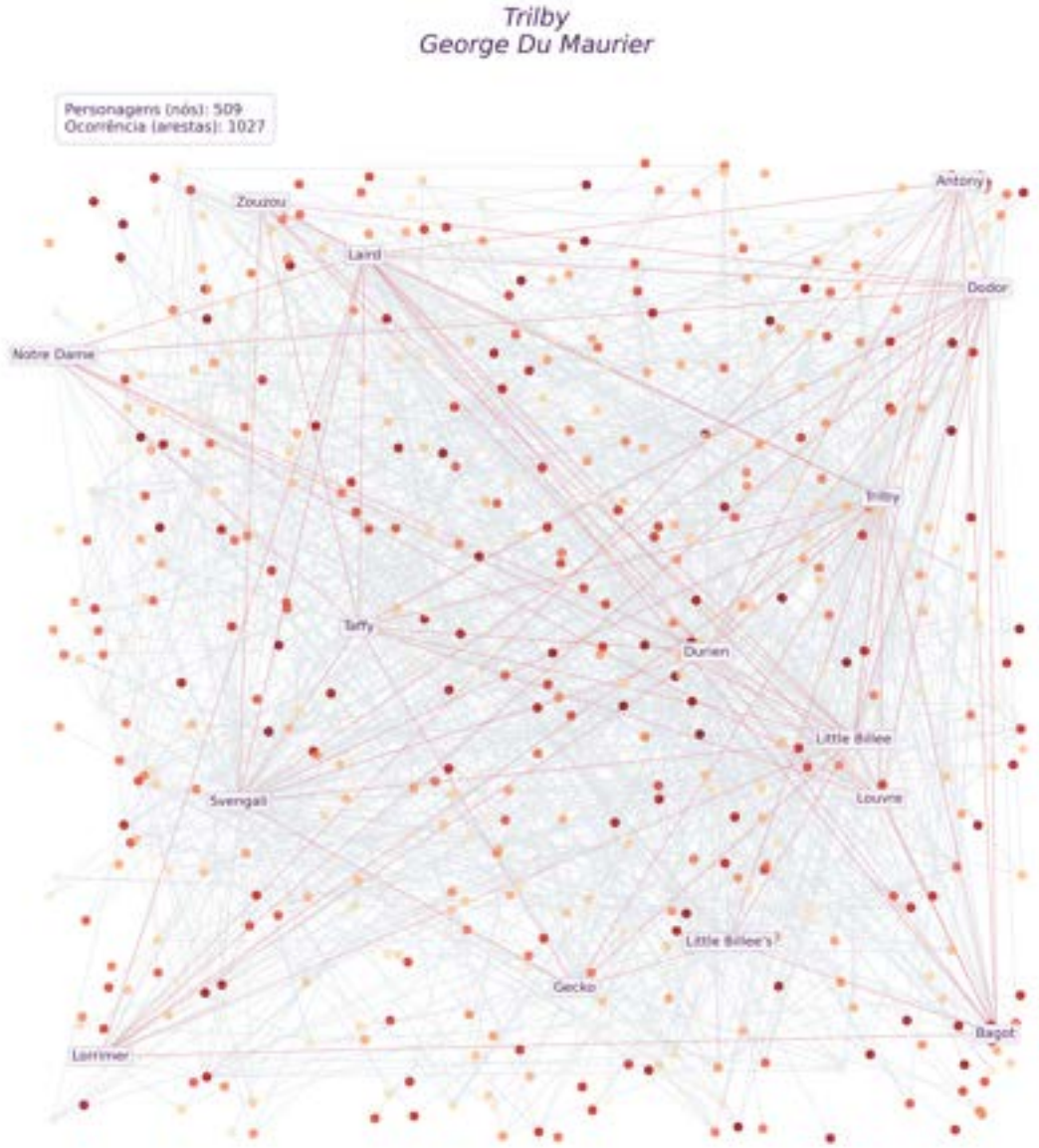

Figura 9 – Rede aleatória construida para o livro: *Trilby*, pertencente à classificação "sucesso".

Figura 10 – Rede aleatória construida para o livro: *The Adventures of Captain Horn*, pertencente à classificação "sucesso".

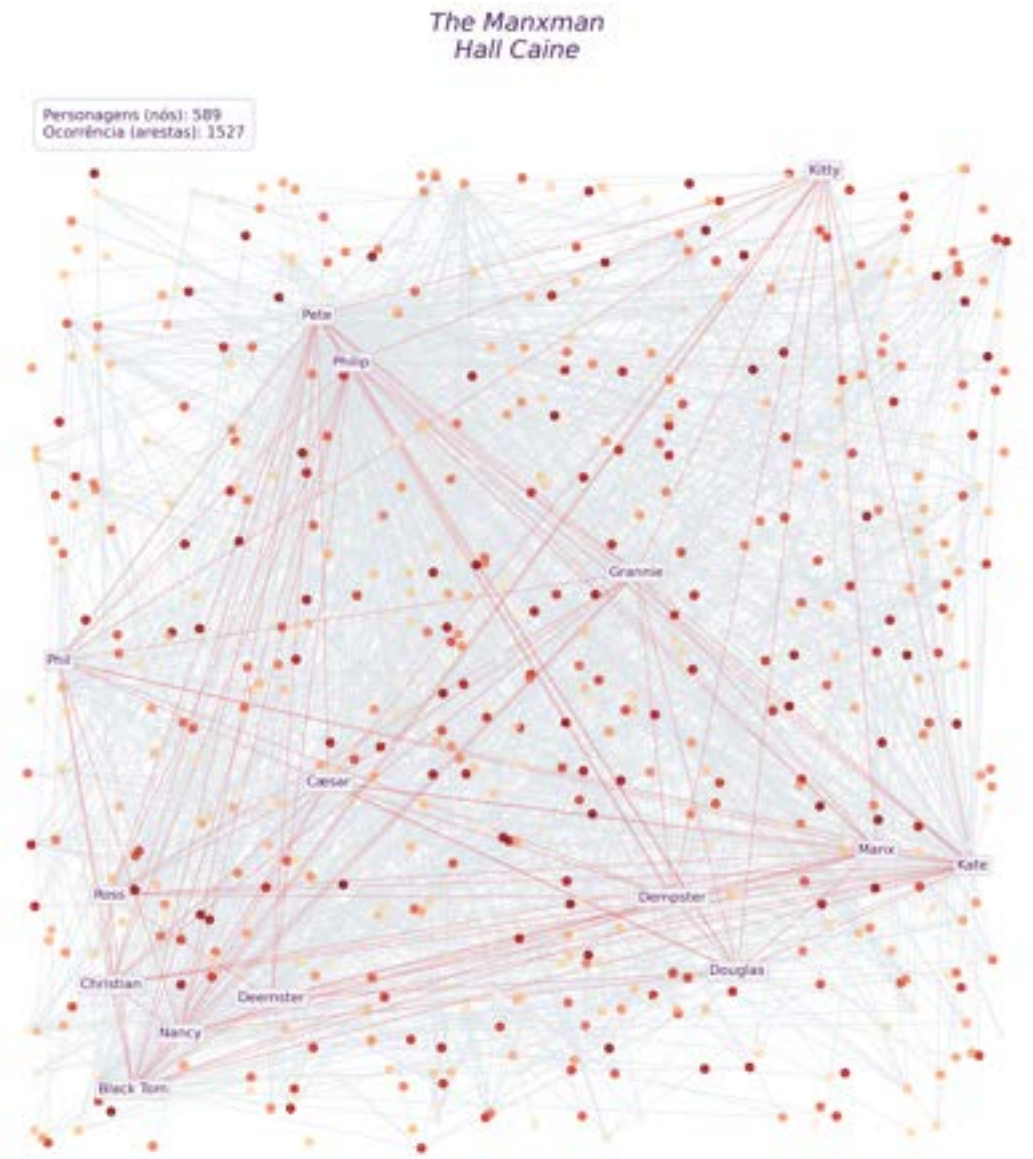

Figura 11 – Rede aleatória construida para o livro: *The Manxman*, pertencente à classificação "sucesso".

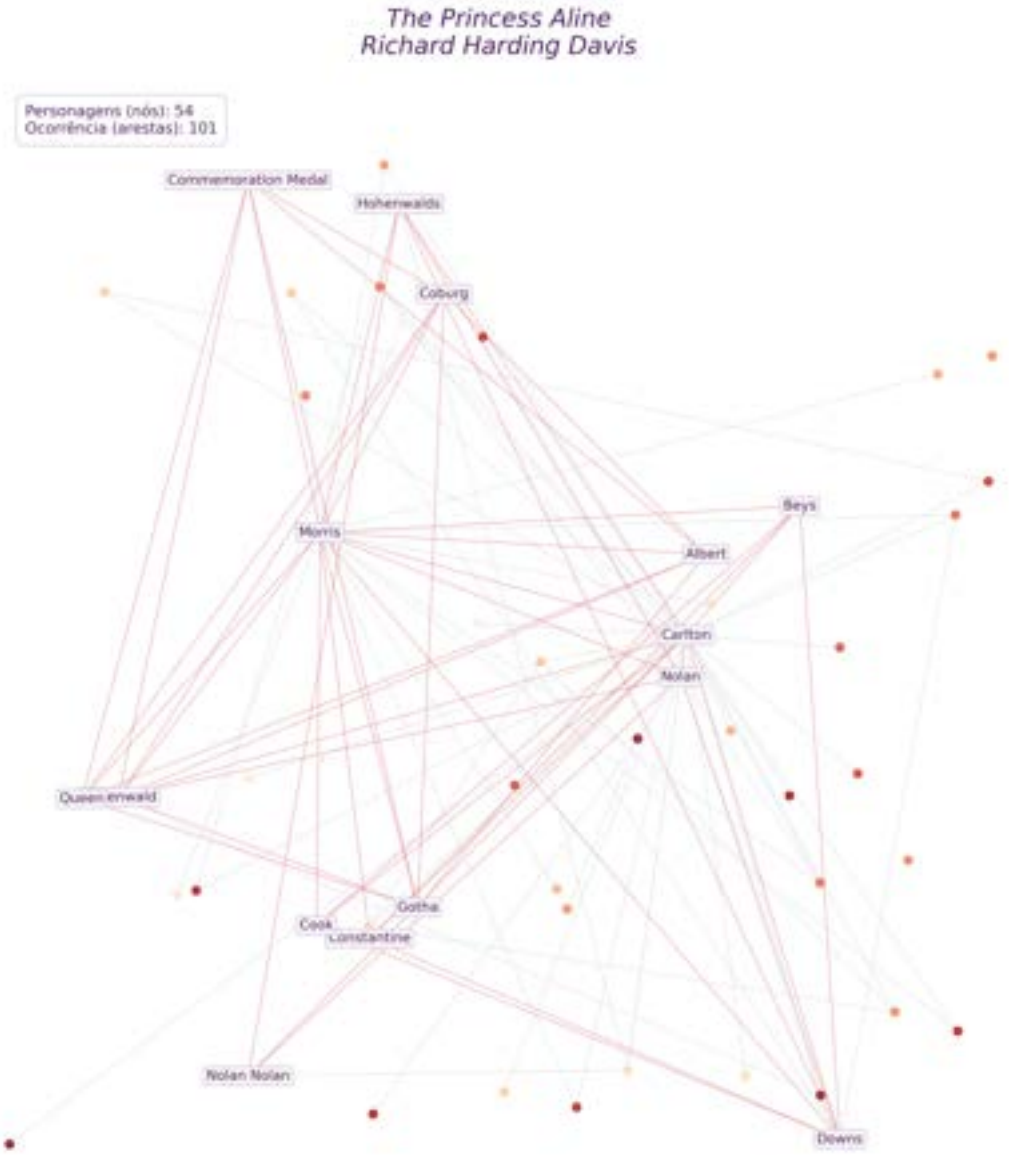

Figura 12 – Rede aleatória construida para o livro: *The Princess Aline*, pertencente à classificação "sucesso".

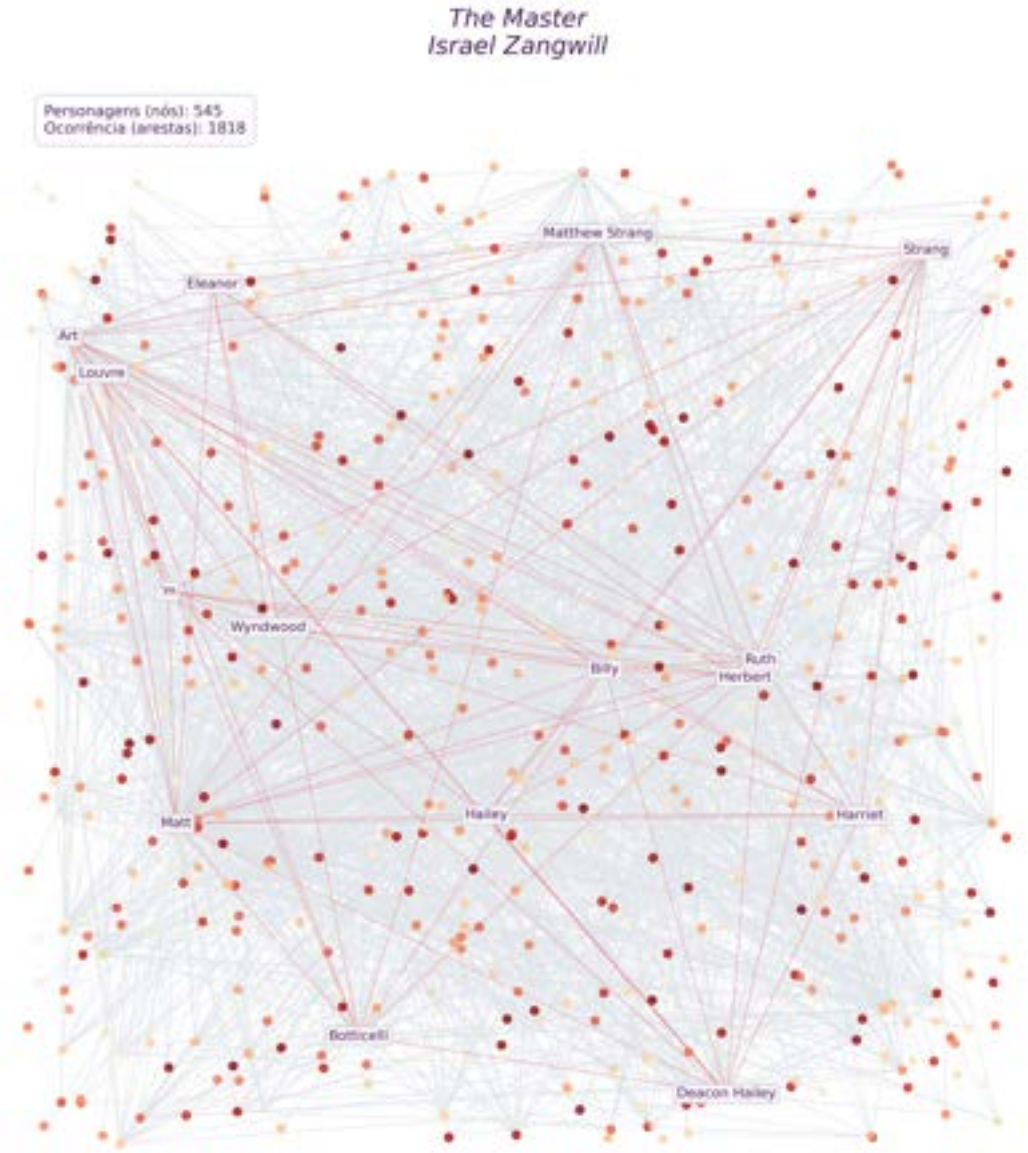

Figura 13 – Rede aleatória construida para o livro: *The Master*, pertencente à classificação "sucesso".

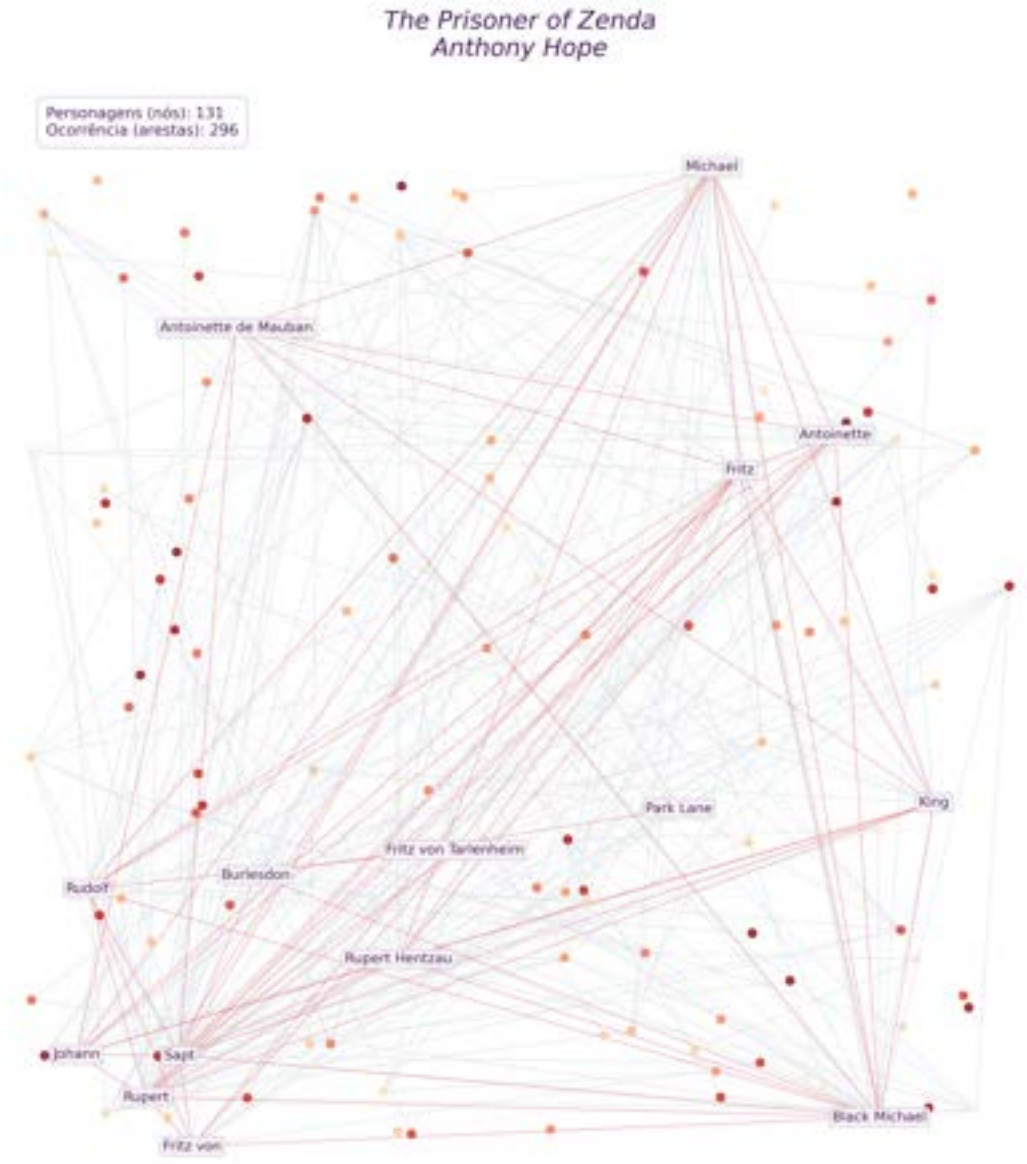

Figura 14 – Rede aleatória construida para o livro: *The Prisoner of Zenda*, pertencente à classificação "sucesso".

Figura 15 – Rede aleatória construida para o livro: *Degeneration*, pertencente à classificação "sucesso".

Figura 16 – Rede aleatória construida para o livro: *My Lady Nobody*, pertencente à classificação "sucesso".

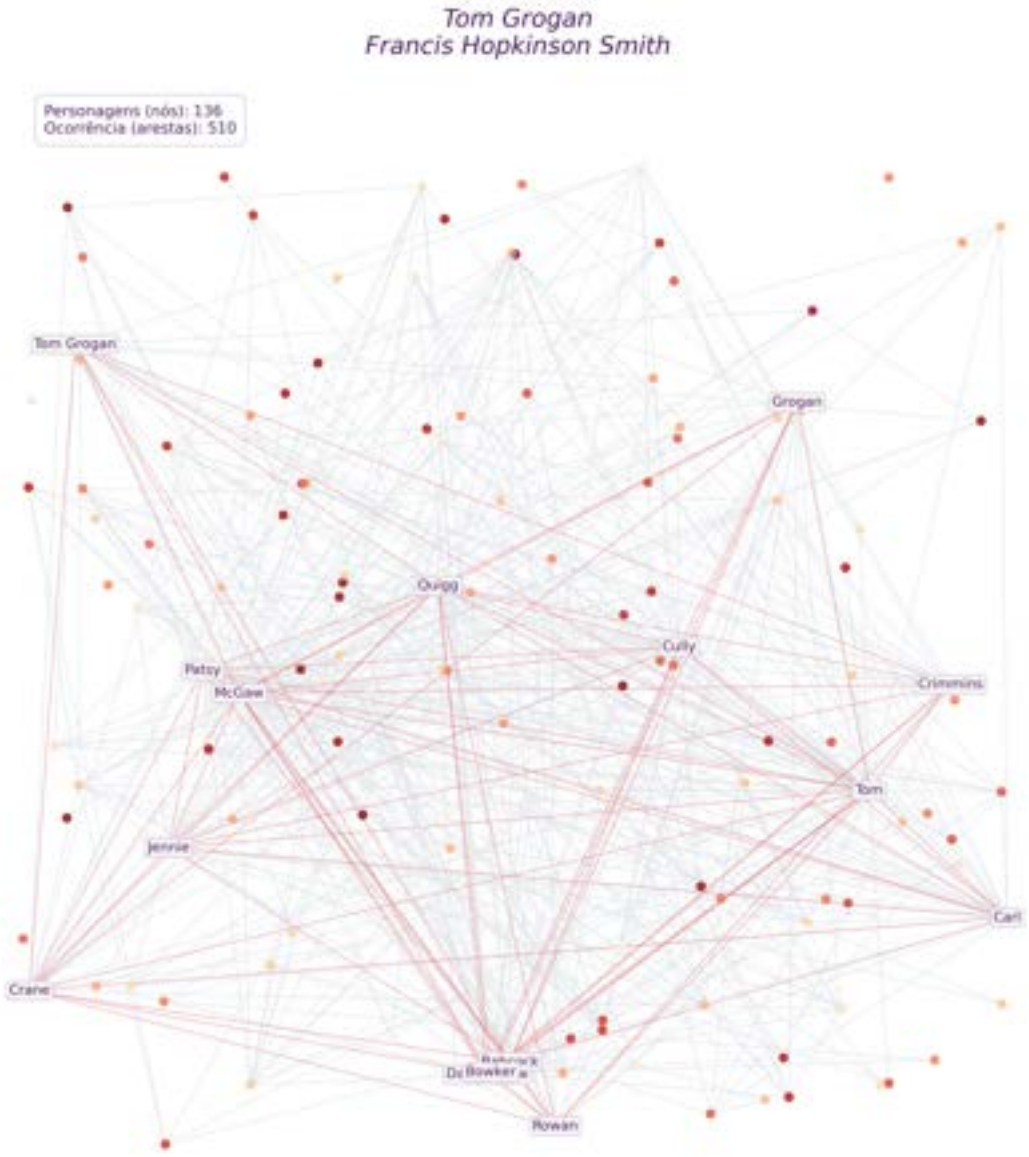

Figura 17 – Rede aleatória construida para o livro: *Tom Grogan*, pertencente à classificação "sucesso".

APÊNDICE C – GRÁFICOS: LIVROS CLASSIFICAÇÃO "NÃO SUCESSO"

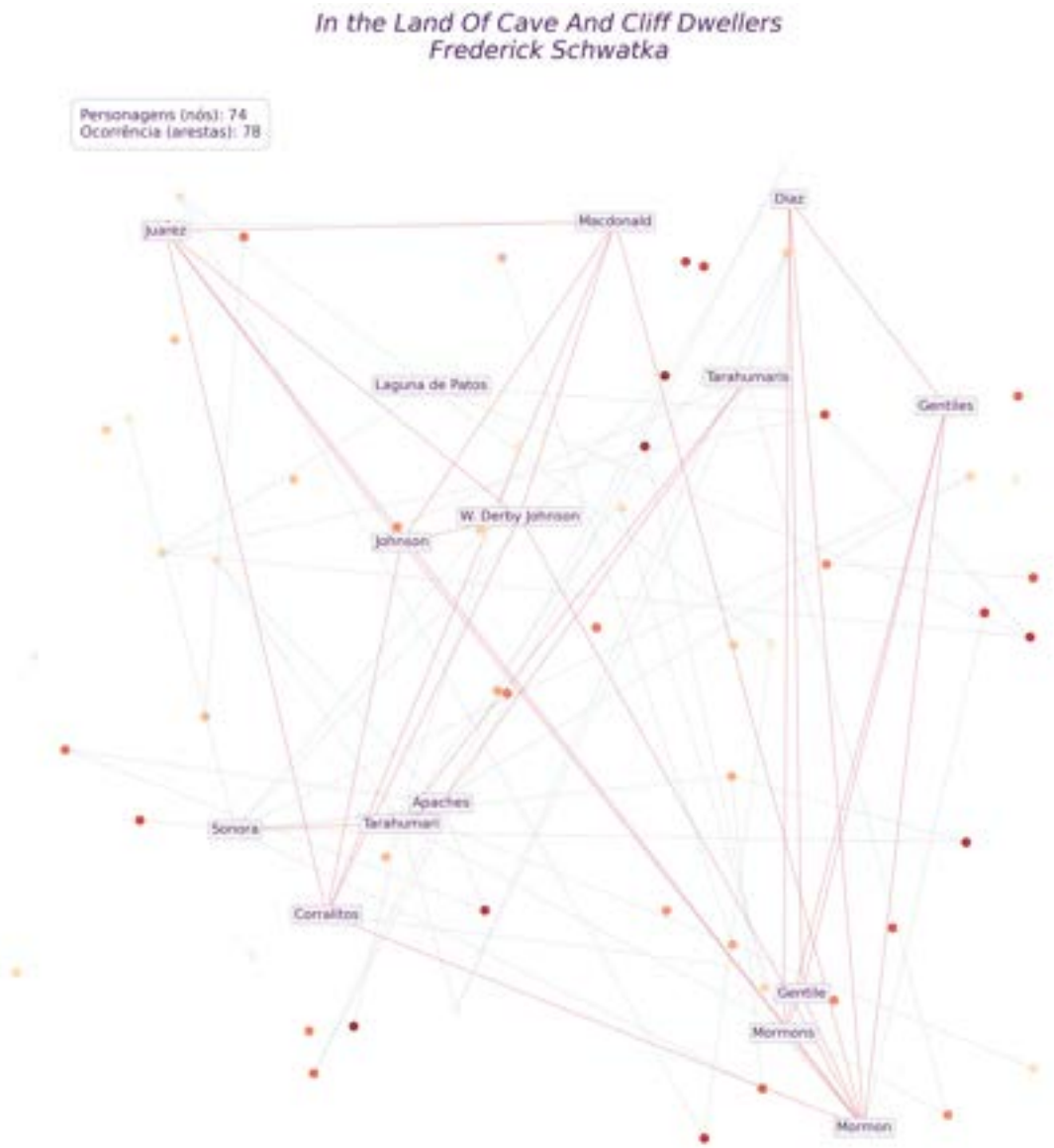

Figura 18 – Rede aleatória construida para o livro: *In the Land of Cave and Cliff Dwellers*, pertencente à classificação "não sucesso".

Figura 19 – Rede aleatória construida para o livro: *Jude the Obscure*, pertencente à classificação "não sucesso".

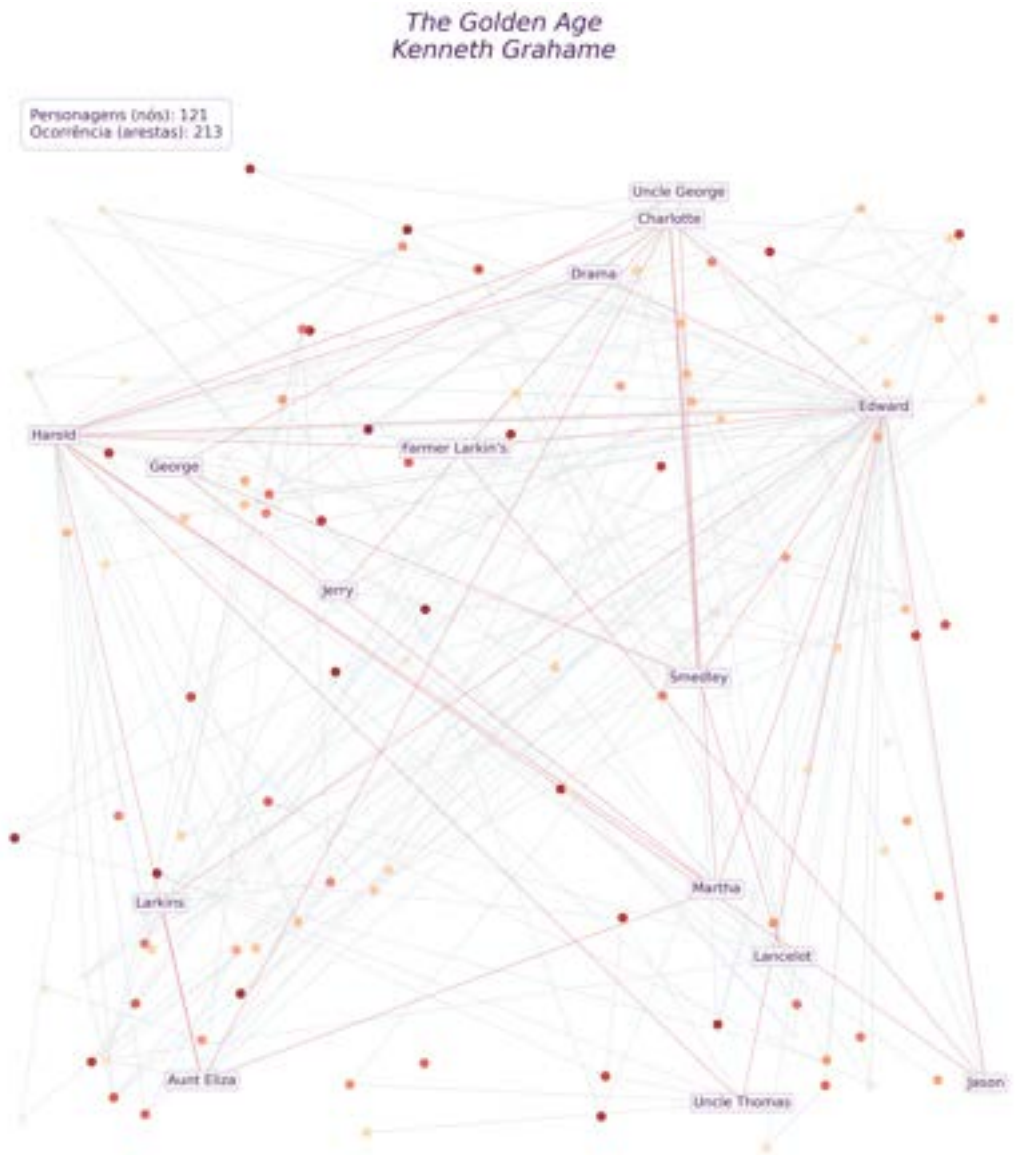

Figura 20 – Rede aleatória construida para o livro: *The Golden Age*, pertencente à classificação "não sucesso".

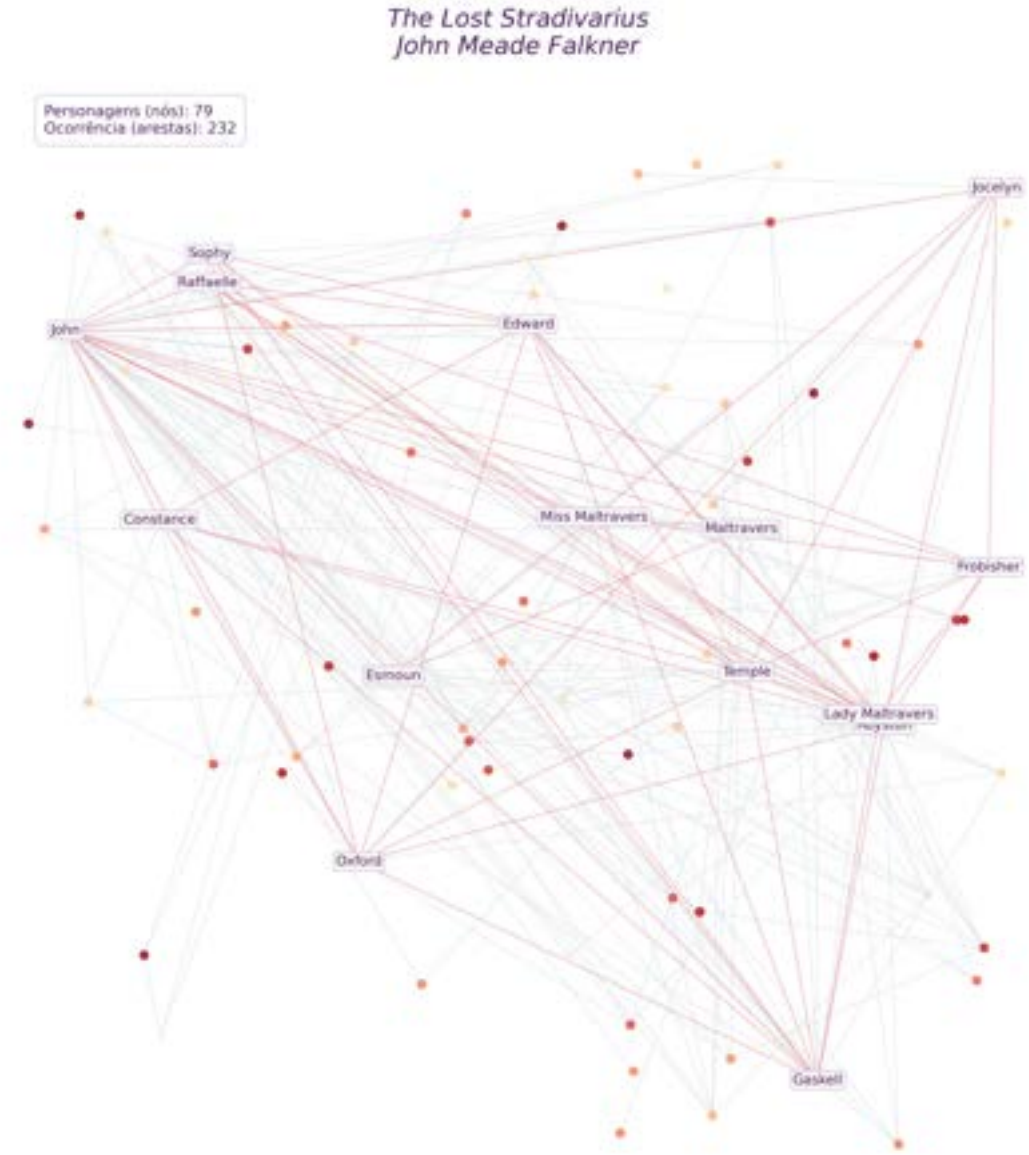

Figura 21 – Rede aleatória construida para o livro: *The Lost Stradivarius*, pertencente à classificação "não sucesso".

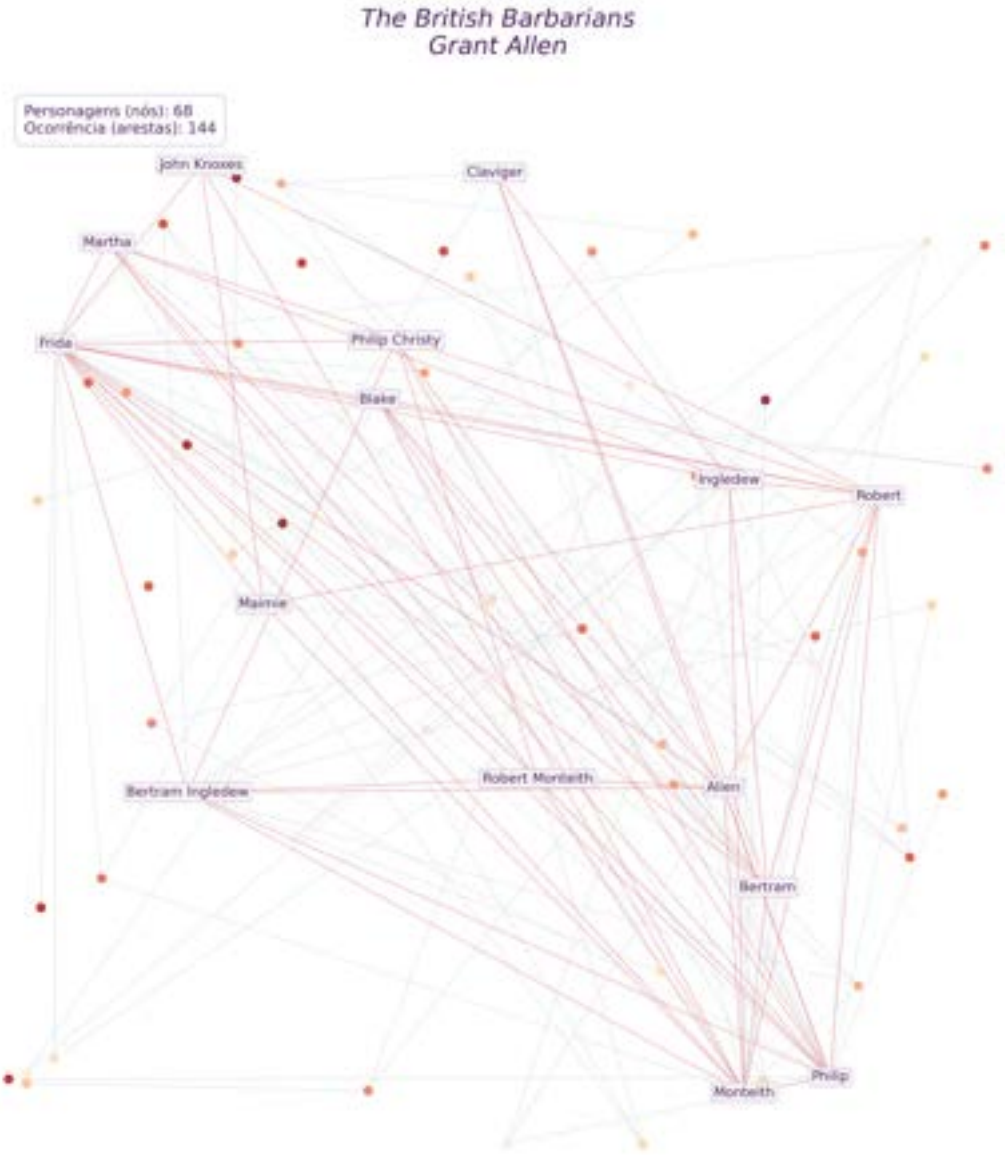

Figura 22 – Rede aleatória construida para o livro: *The British Barbarians*, pertencente à classificação "não sucesso".

Figura 23 – Rede aleatória construida para o livro: *The Sorrows of Satan*, pertencente à classificação "não sucesso".

Figura 24 – Rede aleatória construida para o livro: *Rose of Dutcher's Coolly*, pertencente à classificação "não sucesso".

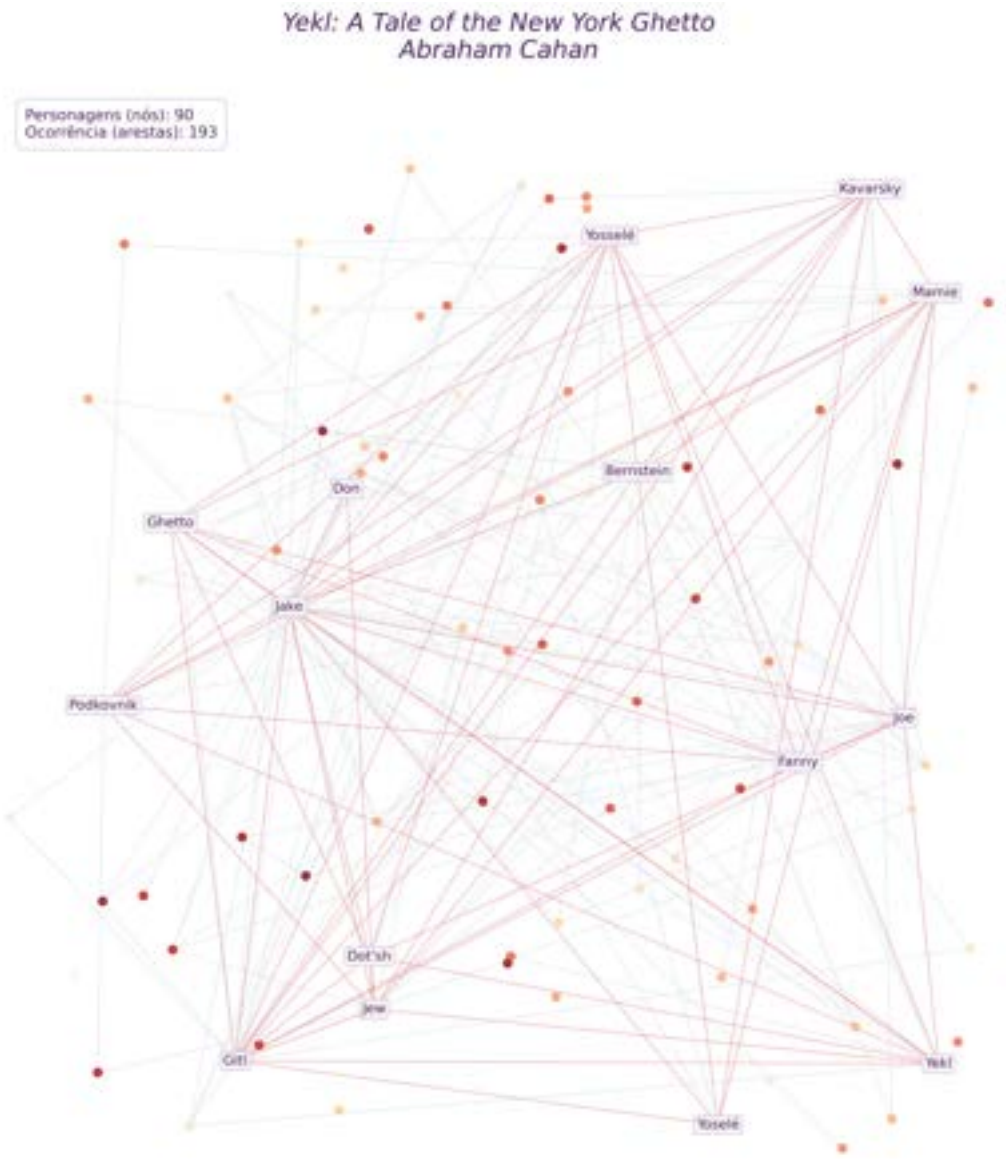

Figura 25 – Rede aleatória construida para o livro: *Yekl: A Tale of the New York Ghetto*, pertencente à classificação "não sucesso".

Figura 26 – Rede aleatória construida para o livro: *Madelon*, pertencente à classificação "não sucesso".

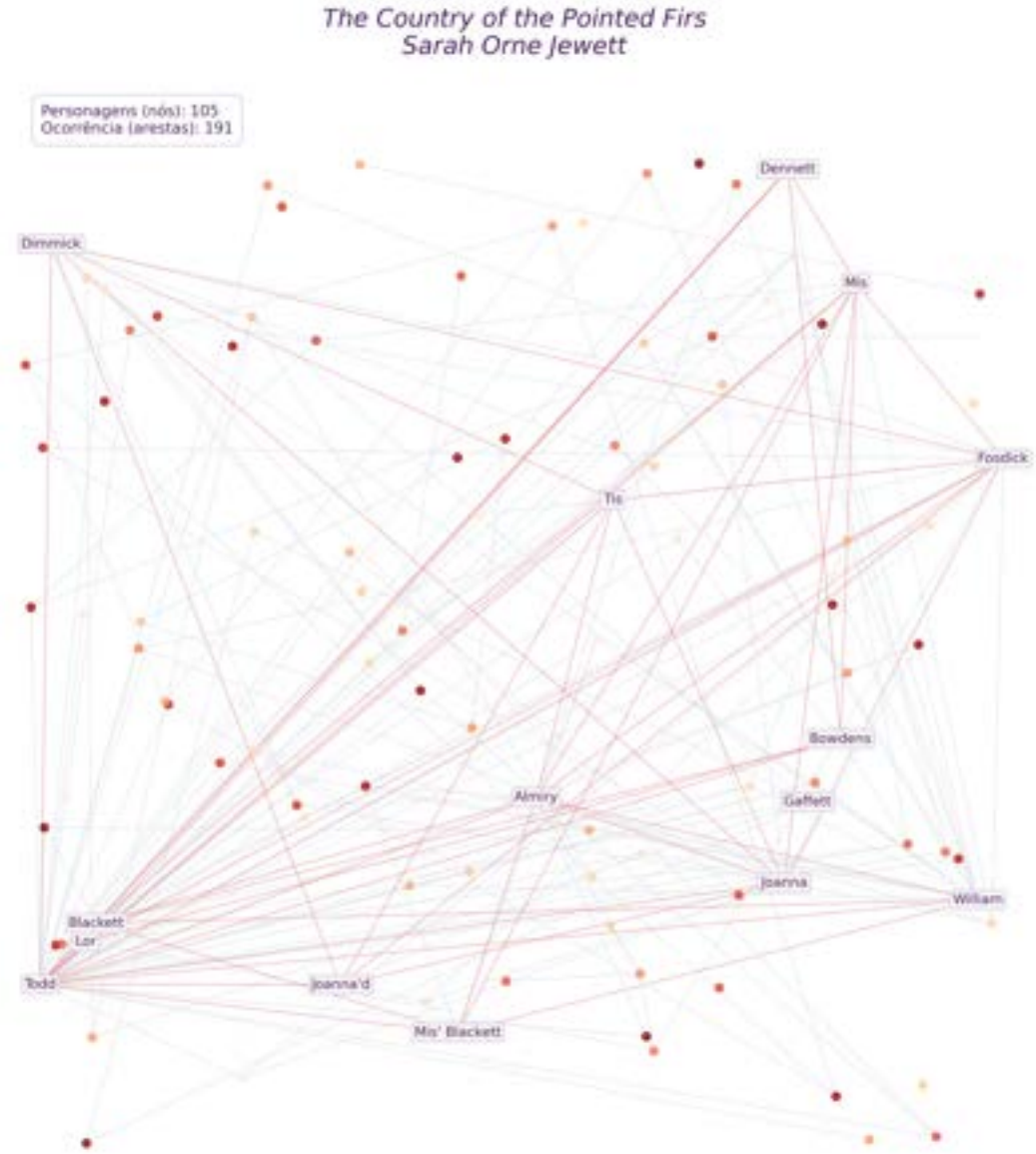

Figura 27 – Rede aleatória construida para o livro: *The Country of the Pointed Firs*, pertencente à classificação "não sucesso".