Universidade Federal do ABC

Seleção de avaliadores de projetos: uma abordagem baseada em PLN e algoritmos de emparelhamento máximo

Projeto de Graduação em Computação III

Aluna
Amanda Cruz Francesconi
RA: 11020315
amanda.cruz@aluno.ufabc.edu.br

Orientador
Prof. Dr. Jesús P. Mena-Chalco
jesus.mena@ufabc.edu.br

14 de junho de 2022

Resumo

Um processo imprescindível ao se realizar pesquisas científicas em geral é a avaliação por pares (*peer review*), onde o projeto ou artigo desenvolvido deve ser avaliado por ao menos um pesquisador qualificado que tenha conhecimento da área de estudo em questão. Visando resolver o problema de atribuição de avaliadores para projetos de pesquisa no âmbito acadêmico – uma vez que, atualmente, esse processo é feito de forma manual e onerosa – são estudadas neste projeto heurísticas computacionais de Processamento de Linguagem Natural para a criação de um grafo bipartido ponderado não direcional de afinidade entre avaliadores e o projeto avaliado.

Adicionalmente é utilizado um algoritmo de emparelhamento de peso máximo para a atribuição/seleção de avaliadores para projetos, considerando que o peso das arestas é correspondente a um coeficiente de correlação entre os dois vértices adjacentes. O coeficiente, por sua vez, é criado a partir de uma análise de dois conjuntos: (1) as palavras que compõem título, resumo e palavras-chave do projeto a ser avaliado e (2) as palavras de títulos e palavras-chave de artigos publicados em revistas e congressos pelo possível avaliador.

Foi possível realizar diversos testes com dados reais na ferramenta construída e obteve-se resultados satisfatórios de emparelhamento entre projeto e avaliador.A obtenção do grafo com melhor resultado de emparelhamento máximo foi alcançado ao se utilizar as técnicas de processamento de linguagem natural (PLN) e utilizando um modelo de saco-de-palavras somente com palavras e não bigramas. Além disso a aplicação desenvolvida foi utilizada para um caso real de auxílio na atribuição de avaliadores de uma instituição de pesquisa. Ainda, uma vez construído, o grafo de projetos e possíveis avaliadores pode ser utilizado para outros fins, como descobrir a correlação entre dois pesquisadores, ferramenta de busca de trabalhos com mesmo tema, identificação de principais áreas de estudo em uma instituição de ensino, entre outras. Posto isto, esse projeto se mostra relevante para a área uma vez que não foi encontrado na literatura uma solução semelhante para o problema proposto.

Palavras-chave: Problema de atribuição, grafo bipartido, emparelhamento de peso máximo, processamento de linguagem natural.

Abstract

An essential process when carrying out scientific research in general is peer review, where the project or article developed must be evaluated by at least one qualified researcher who has knowledge of the area of study in question. In order to solve the problem of assigning evaluators to research projects in the academic field – since, currently, this process is done manually and in an onerous way – computational heuristics of Natural Language Processing is studied in this project to create a non-directional weighted bipartite graph of affinity between evaluators and the evaluated project.

Additionally, a maximum weight matching algorithm is used for the assignment/selection of evaluators for projects, considering the weight of the edges correspond to a correlation coefficient between the two adjacent vertices. The coefficient, in turn, will be created from an analysis of two sets: (1) the words that compose the title, abstract and keywords of the project to be evaluated and (2) the words of titles and words of articles published in journals and conferences by the potential evaluator, data that were captured from the Lattes platform.

The expected result was obtained, with the tool generating a coherent pairing between projects and evaluators. Obtaining the final graph and its maximum matching was achieved by using natural language processing (NLP) techniques and using the bag-of-words only with words and not bigrams. In addition, the developed application was used for a real case in the attribution of evaluators of a research institution. Also, once built, the graph of projects and possible evaluators can be used for other purposes, such as discovering the correlation between two researchers, a tool to search for works with the same theme, identification of main areas of study in an educational institution, among others. Hereupon, this project is relevant to the area since a similar solution to the proposed problem was not found in the literature.

Keywords: Assignment problem, bipartite graph, maximum weight matching, natural language processing.

Agradecimentos

Gostaria de agradecer primeiramente a Universidade Federal do ABC onde obtive um ensino de excelência. Agradeço ao meu orientador Prof. Dr. Jesús P. Mena-Chalco por todo auxílio nesse projeto e por todos os ensinamentos passados durante os anos de Universidade. E ao meu namorado Rafael Bruno Ferreira Figueiredo por todo companheirismo e apoio.

Sumário

1 Introdução

A pesquisa científica é essencial tanto no Brasil como no mundo e se faz cada vez mais necessária. Dada a sua importância, a pesquisa deve ser incentivada, disseminada e seus processos facilitados, dentro do possível (Zhang *et al.*, 2022). Uma das etapas da pesquisa científica é a revisão por pares.

A tarefa de revisão por pares é realizada em diferentes âmbitos, desde dentro de uma universidade, onde é necessário a escolha de avaliadores (que serão professores da própria universidade) para projetos em editais internos em que irão se decidir quais os melhores projetos para ganho de bolsas, até no âmbito nacional em órgãos como CAPES e CNPq, onde pesquisadores do Brasil inteiro podem ser possíveis avaliadores e um número representativo de projetos são enviados anualmente para que seja feita essa análise.

Outro contexto em que se é necessário fazer a escolha de avaliadores é nas revistas científicas, onde os editores são responsáveis por atribuir pesquisadores a artigos que devem ser avaliados e futuramente publicados na revista, de forma que esses profissionais têm um papel fundamental para a qualidade da produção científica (Silva C., 2016). Porém, cada vez mais com a especialização dos pesquisadores em assuntos determinados, poucos são capazes de julgar de forma justa e coerente cada um dos assuntos (Terán, 2011). Por esse motivo é de extrema importância a correta escolha de avaliadores para os artigos, a fim de garantir que o avaliador terá pleno entendimento do assunto para avaliá-lo (Checco, 2021).

A proposta desse trabalho é utilizar diversas áreas de conhecimento combinadas para construir um sistema que tem como objetivo auxiliar em uma das etapas do processo de construção da pesquisa científica no país, que é a escolha assertiva de pesquisadores para avaliar projetos de pesquisa (*peer review*) (Forsberg, 2022). Isso é realizado da seguinte forma, a partir de currículos de avaliadores (e.g., CVs Lattes) e um conjunto de projetos a serem avaliados, encontrar o melhor pareamento avaliador-projeto. Considerando os conhecimentos do avaliador e o tema abordado no projeto. As áreas envolvidas são teoria dos grafos, processamento de linguagem natural, algoritmos de emparelhamento máximo, todas descritas com maior detalhamento na Seção 3 (conceitos básicos).

Dois estudos que se propõem a resolver problemas na mesma área são o trabalho desenvolvido por Mrowinski (2017) que utiliza algoritmos evolutivos bio-inspirados para auxiliar na redução do tempo necessário para realizar a avaliação por pares em um contexto de revistas científicas, determinando quantas solicitações de revisão devem ser enviadas e em qual momento, buscando a otimização desse processo; e o artigo de Checco (2021), que utiliza uma rede neural para tentar prever qual nota os avaliadores darão a um determinado artigo, porém ressalta que devem ser observadas possíveis tendências

na rede assim como problemas éticos sobre esse método de revisão por inteligência artificial. O trabalho cita que as técnicas de aprendizado de máquina são compostas por algoritmos conservadores, logo isso poderia levar a uma propagação de valores e costumes já existentes nas avaliações existentes hoje. Além disso, o algoritmo poderia beneficiar certos autores ou países considerando o histórico de boas avaliações enquanto, por outro lado, regiões com baixa representatividade em publicações podem ter suas avaliações subestimadas pelo mesmo motivo. Outro ponto citado é a questão da transparência. Um autor, ao ter sua nota, deve poder questionar o motivo desta e atualmente os modelos automáticos propostos não retornam o racional que levou a determinada nota.

2 Objetivos

O objetivo desse trabalho é resolver o problema de determinação de avaliadores em uma revisão por pares através da criação de um sistema, em que, a partir de um conjunto de avaliadores com currículo (e.g. CV Lattes) e um conjunto de projetos de pesquisa a serem avaliados, utilizando um algoritmo de emparelhamento máximo, seja determinado quais projetos serão designados para cada avaliador.

2.1 Objetivos específicos

Durante a realização do trabalho, alguns objetivos auxiliares também são considerados:

- Tratamento correto dos dados de entrada e análise das melhores abordagem de mineração de dados para facilitar a captura das informações relevantes nos currículos dos avaliadores e nos projetos submetidos.
- Criação do grafo bipartido e análise das correlações geradas pelos algoritmos a serem utilizados.
- Análise do emparelhamento final gerado pelo processo construído afim de verificar se resultado é coerente e utilizável.

3 Conceitos básicos

Nesta seção apresentam-se os conceitos básicos necessários para o desenvolvimento e entendimento do trabalho realizado. Iniciando pelo conceito de revisão por pares, que é o grande tema retratado nesse trabalho, posteriormente o PLN que é utilizado no início do método para refinamento dos dados de entrada, em sequência conceitos de grafos que são estruturas fundamentais para a resolução que será proposta e por fim o conceito do emparelhamento máximo, método que é utilizado para resolver o problema proposto.

3.1 Revisão por pares

A **revisão por pares** é o procedimento utilizado no meio científico para avaliação de trabalhos para um fim, seja, por exemplo, para recebimento de bolsas acadêmicas ou para determinação de artigos a serem publicados em uma revista, onde o trabalho é avaliado por outros pesquisadores que tenham conhecimento sobre o assunto. Ela é utilizada para "avaliar o valor de uma proposta pela sua contribuição para o avanço do conhecimento", como indica Serra F. (2007). Normalmente os critérios considerados são originalidade, importância, relevância do assunto, qualidade da argumentação e a linguagem utilizada (Checco, 2021).

Porém esse método sofre algumas críticas por parte dos pesquisadores como: o sistema está sujeito a injustiças, as normas de avaliação não são claras, o processo não está imune a idiossincrasias (Conix, 2021), há abusos e preconceitos por parte de editores e revisores, sua lentidão provoca atrasos na publicação, o processo é pobre na detecção de erros, é quase inútil na detecção de fraudes e má conduta e o processo é caro e trabalhoso (Patrus R., 2016).

3.2 Processamento Linguagem Natural

O **processamento de linguagem natural** é uma área que envolve ciência da computação, inteligência artificial e linguística e pode ser definido como a seguir:

> O processamento de linguagem natural é uma área de pesquisa em ciência da computação e inteligência artificial (IA) preocupada com o processamento de linguagens naturais como o inglês ou o mandarim. Esse processamento geralmente envolve a tradução da linguagem natural em dados que um computador pode usar para aprender sobre o mundo. E essa compreensão do mundo às vezes é usada para gerar texto em linguagem natural que reflete essa compreensão. (Lane, 2019)

Por esse processamento envolver a linguagem humana, existem algumas dificuldades inerentes pois a linguagem é aprendida intuitivamente. O PLN envolve diversas técnicas com o intuito de mitigar essa questão e algumas delas são utilizadas nesse projeto e serão explicadas a seguir (Eisenstein, 2019).

O PLN é essencial para o desenvolvimento desse projeto, uma vez que é realizado um tratamento nas informações contidas nos currículos Lattes dos avaliadores envolvidos, assim como no projeto a ser avaliado. Para isso utiliza-se os conceitos de radicalização, bigramas e modelo saco-de-palavras.

3.2.1 Radicalização

A a **radicalização** (*stemming*) é uma técnica de processamento textual para obtenção de informação que consiste em reduzir palavras flexionadas à sua base (ou tronco, do inglês *stem*) (Orengo, 2001), ou seja, de simplificar o conteúdo presente em cada palavra. Por exemplo as palavras *estudo, estudante, estudar, estudaria* seriam reduzidas ao radical *estud* que já traz em si o significado necessário que todas essas palavras representam, apesar das diversas variações. O radical normalmente não será uma palavra gramaticalmente correta, como no caso desse exemplo, porém tem o valor necessário da palavra original, servindo bem ao propósito de análises (Alves, 2021).

Esse processo, já utilizado na área da linguística e agora aplicado ao processamento de linguagem natural, é essencial para o projeto desenvolvido, uma vez que as palavras são comparadas e o importante é a comparação entre os significados e não as diversas variações que a palavra pode ter.

3.2.2 Bigramas

Uma análise por n-grama é uma técnica utilizada no processamento textual onde se forma uma sequência de n itens dentro de uma frase. Bigramas por sua vez, são n-gramas com $n = 2$, ou seja, são essencialmente sequências de duas palavras (Hachaj T., 2018; Jurafsky & Martin, 2000). Esse modelo é utilizado para facilitar a análise de textos

em linguagem natural. Como exemplo, na frase “como avaliar esse trabalho?”, temos os bigramas: “como avaliar”, “avaliar esse” e “esse trabalho”.

Esse será um dos conceitos utilizados no PLN dos currículos Lattes e nos projetos, para identificação da quantidade de bigramas em comum entre ambos, definindo assim a similaridade.

A partir dos bigramas construídos é utilizado um modelo de saco-de-palavras que ignora a posição desses bigramas no texto, importando-se somente com o conteúdo e quantidade de vezes em que esses bigramas aparecem no texto de referência (Eisenstein, 2019).

3.2.3 Modelo saco-de-palavras

Também conhecido pelo termo em inglês *bag-of-words* o **modelo saco-de-palavras** é uma técnica de processamento textual onde um texto é simplificado e representado por um conjunto não ordenado de palavras como no exemplo da Figura 1 (Qader W., 2019).

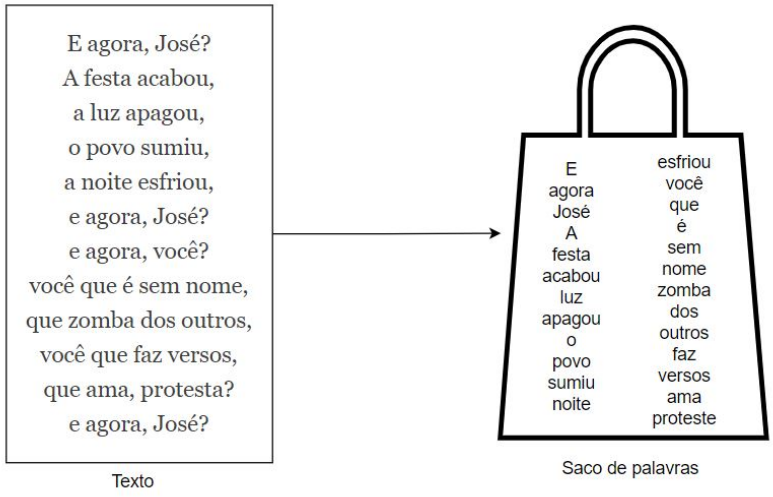

Figura 1: *Exemplo de um saco-de-palavras.*

Dessa forma, temos como representar os dados contidos em um texto e utilizá-los como entrada para os próximos passos. Nesse modelo são desconsiderados estrutura gramatical e ordem das palavras, dando importância somente para o conteúdo das palavras contidas no texto.

3.2.4 Stop words

As **stop words**, ou palavras vazias, são consideradas palavras que não carregam significado por elas mesmas Rajaraman (2014). As *stop words* são as palavras mais comuns

em um idioma assim como preposições e artigos. Temos como exemplo no português *que, quando, isso, antes, lhe, uma*, etc. Essas palavras costumam ser retiradas antes de serem realizadas analises textuais para que os modelos ou classificações sejam mais precisos.

3.3 Conceitos de grafos

Para a realização deste trabalho é essencial o entendimento de conceitos em grafos.

Um grafo G é um par ordenado *(V(G), E(G))* que consiste em um conjunto *V(G)* de vértices e um conjunto *E(G)* de arestas com uma função de incidência ψ_G que associa cada aresta de G com um par não ordenado de vértices de G (Bondy & Murty, 1976).

Os vértices são usados para representar os possíveis avaliadores e projetos a serem avaliados. As arestas indicarão a afinidade entre avaliador e o projeto e o peso das arestas é calculados a partir de um coeficiente de similaridade. Logo, é utilizado um grafo bipartido ponderado, cujas definições estão a seguir.

3.3.1 Grafos bipartidos

Um tipo específico de grafo são os **grafos bipartidos**. Um grafo é bipartido quando é possível dividir o conjunto de vértices em duas partes X e Y de modo que toda aresta conecte um vértice de X com um de Y.Um grafo G com bipartições (X,Y) é denotado por $G[X, Y]$ (Bondy & Murty, 1976).

Esse tipo de grafo é útil neste trabalho uma vez que temos a parte dos avaliadores e outra dos projetos a serem avaliados. Uma aresta deve, portanto, conectar um projeto a um avaliador (conectando dois vértices de partes diferentes) como exemplificado na Figura 2.

3.3.2 Grafos ponderados

Grafos ponderados são aqueles em que cada aresta recebe um valor que representa o relacionamento entre os vértices. Dado um grafo G, para cada aresta e de G atribui-se um número real $w(e)$ chamado peso. Assim, G em conjunto com os pesos em suas arestas é denominado **grafo ponderado** e denotado por (G, w) (Bondy & Murty (1976)).

Por exemplo, em um grafo onde os pesquisadores são os vértices, o valor da aresta pode representar a quantidade de artigos publicados entre ambos, como representado na Figura 3.

Neste trabalho, o valor atribuído a uma aresta é a afinidade entre avaliador e projeto avaliado considerando um coeficiente de similaridade desenvolvido para esse fim. Ainda

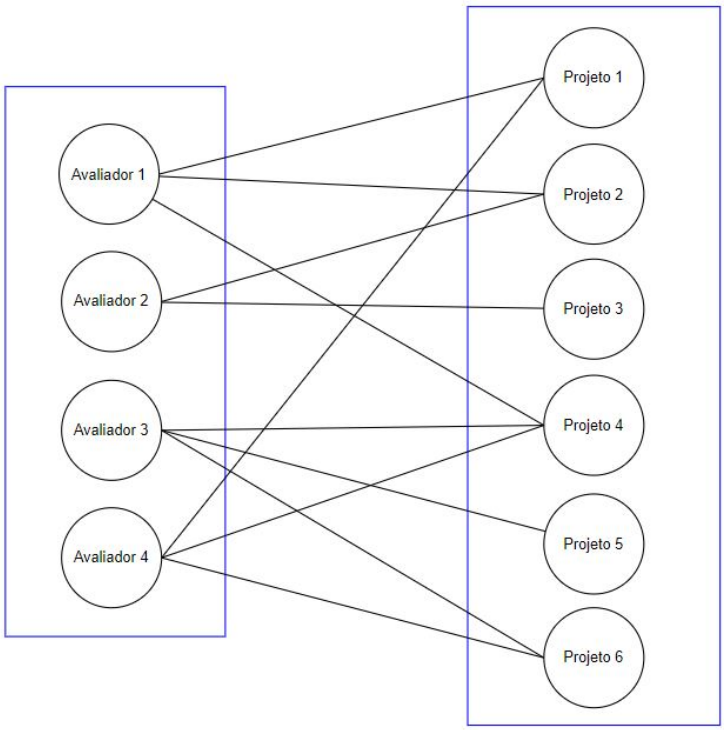

Figura 2: *Exemplo de um grafo bipartido.*

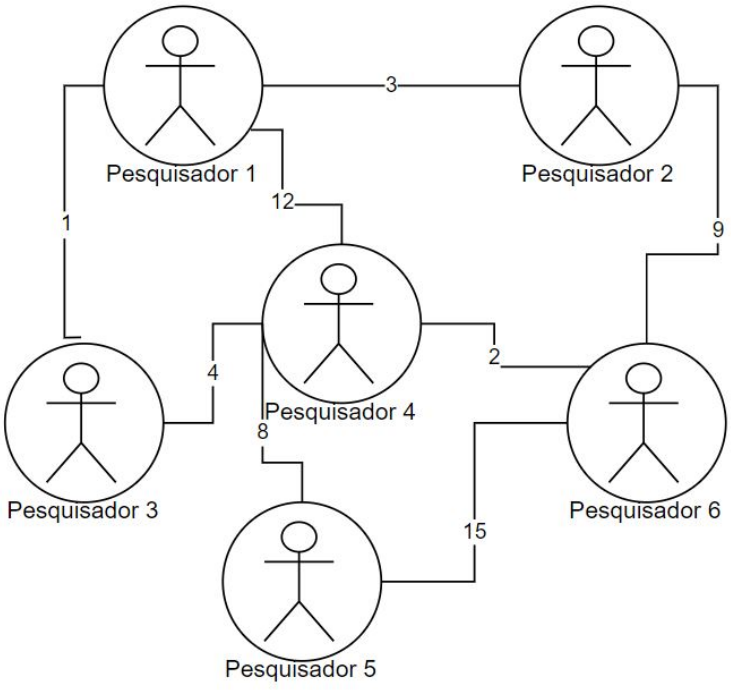

Figura 3: *Exemplo de um grafo ponderado.*

considerando um grafo bipartido onde só existem arestas entre avaliador e projeto.

3.3.3 Caminhos

Em um grafo, um **caminho** é uma sequência alternada de vértices e arestas em que tanto os vértices quanto arestas são distintos, ou seja, não se repetem (Prestes, 2020). Sendo v são vértices e a arestas, um caminho w pode ser representado como:

$$w = v_0, a_1, v_1, a_2, v_2, ..., a_n, v_n$$

3.4 Emparelhamento máximo

Um emparelhamento M em um grafo $G = (V, A)$ é um subconjunto de arestas $M \subseteq A$ que não compartilham arestas entre si Prestes (2020). Um emparelhamento máximo pode ser definidos como:

> O peso de um emparelhamento é a soma dos pesos das arestas. O **emparelhamento máximo** é aquele em que caso se adicione mais arestas não continuará sendo um emparelhamento. A cardinalidade de um emparelhamento é o número de vértices emparelhados (doccumentation, n.d.).

O conceito de caminho aumentador é essencial para o entendimento do algoritmo para encontrar um emparelhamento de peso máximo. O algoritmo utilizado é resolvido em etapas e em cada etapa temos um emparelhamento M. Inicialmente M é vazio. Um vértice i se diz emparelhado se existe uma aresta (i, j) em M e desemparelhado caso contrário. Um caminho aumentador é um caminho em que as duas extremidades estão em vértices desemparelhados e os vértices interiores estão emparelhados Galil (1986).

Para fins didáticos é descrito abaixo um possível método para se encontrar um emparelhamento máximo em um grafo bipartido e seu pseudocódigo.

O algoritmo é realizado em fases e termina quando não existirem mais caminhos extensores. Em cada fase são realizados os seguintes passos:

- Realiza uma busca em largura começando pelos vértices livres (ainda não emparelhados) para encontrar o conjunto maximal de caminhos extensores de comprimento mínimo.
- Com esse conjunto de caminhos extensores, utilizando uma busca em profundidade encontra um subconjunto de caminhos disjuntos (ou seja, um vértice não pode estar em dois desses caminhos escolhidos).
- Após se obter esses caminhos, eles são incorporados ao emparelhamento.

Algorithm 1: Pseudocódigo algoritmo de emparelhamento máximo

Result: Emparelhamento máximo no grafo G

$M = \emptyset$

while *Existe um caminho M-aumentador P em G* **do**

| $M = (M \backslash E(P) \cup (E(P) \backslash M))$

end

devolve M

O algoritmo de emparelhamento máximo utilizado nesse trabalho se baseia no método de *Blossom* para encontrar caminhos aumentadores e no método *primal-dual* para encontrar o emparelhamento de peso máximo, ambos métodos foram inventador por Jack Edmonds (doccumentation, n.d.). Os métodos podem ser encontrados em Edmonds (1965b) e Edmonds (1965a).

4 Trabalhos Correlatos

O método de avaliação por pares se iniciou com o surgimento de periódicos científicos no século XVII, porém só foi oficializado pela *Royal Society of London* em 1753 quando percebeu-se necessário fazer a minuciosa apuração e triagem dos artigos a serem publicados (Patrus R., 2016). Desde então, em sua grande maioria, a seleção de avaliador para cada projeto é feita manualmente.

Algumas ferramentas já existentes auxiliam o processo inicial de avaliação, conferindo a formatação do artigo, se há plágio e até mesmo a escrita como é discutido por Checco (2021). Existem também trabalhos que citam como melhorar a revisão dos artigos em si, como por exemplo utilizando programação genética cartesiana para realizar as primeiras análises do artigo, como descrito por Mrowinski (2017), e reduzindo em até 30% o tempo da revisão.

Porém, segundo o nosso conhecimento, não há artigos publicados em que se resolva de forma automática a designação de projetos ao avaliador mais capacitado para tal, problema que atualmente consome tempo dos centros de pesquisas de Universidades e entidades científicas.

Em síntese, este trabalho é relevante por ter potencial de auxiliar em diversas aplicações e reduzir o tempo de trabalho manual de diversos profissionais no Brasil e no mundo. Apesar de já existirem algumas soluções propostas, essas não são amplamente utilizadas. Com cada vez mais pesquisa nesse campo e evoluções no problema de emparelhamento de avaliadores e projetos será possível poupar trabalho manual de diversas pessoas, além de possivelmente trazer um resultado mais assertivo para esse emparelhamento. A abordagem tratada neste trabalho especificamente não foi encontrada na literatura, sendo

portanto, uma nova visão de resolução desse problema que permeia a pesquisa acadêmica há algumas décadas.

5 Método

A seção a seguir apresenta cada passo da heurística construída para a resolução do problema proposto: encontrar o melhor emparelhamento entre avaliadores e projetos. Em linhas gerais a proposta encontra-se ilustrada na Figura 4.

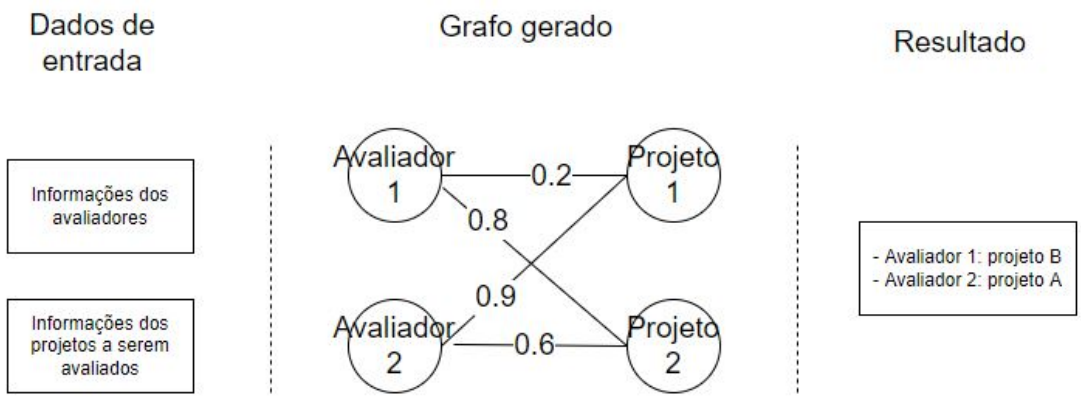

Figura 4: *Fluxo genérico da heurística elaborada.*

5.1 Captura e tratamento dos dados

O passo inicial do projeto foi a captura e tratamento dos dados a serem inseridos no grafo, como está representado na Figura 5. A seguir cada um desses processos é explicado.

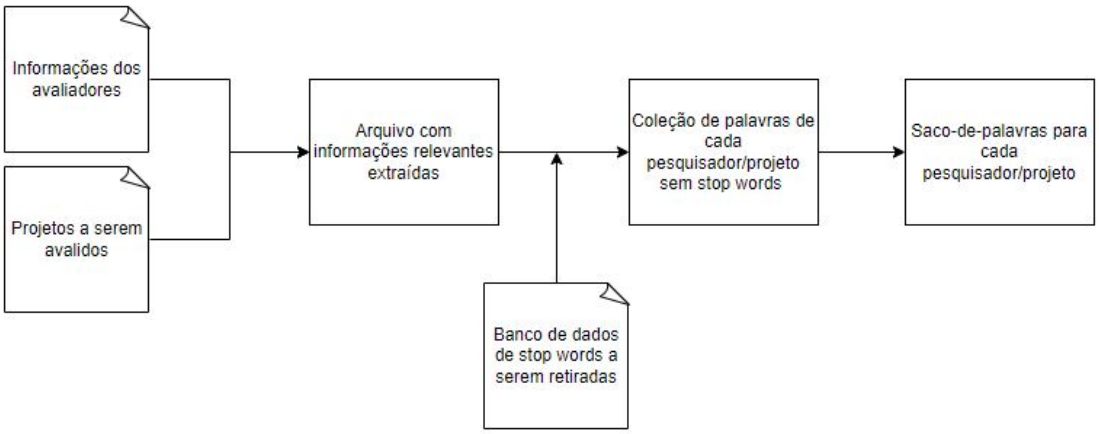

Figura 5: *Diagrama de tratamento de dados do projeto.*

5.1.1 Dados capturados

Os dados utilizados para criação dos vértices de avaliadores são os professores doutores da Universidade Federal do ABC, mais especificamente do Centro de Matemática, Computação e Cognição (CMCC). Esses dados foram coletados em formato de texto da plataforma Lattes pelos membros do time de pesquisa para trabalhos prévios (Mena-Chalco, 2012). Por outro lado, os projetos são amostras realizadas manualmente de projetos reais de alunos.

Os dados capturados de cada uma das partes encontra-se a seguir:

Dados dos avaliadores: • Nome • Título de livros publicados • Título de livros onde publicou um ou mais capítulos • Título e palavras-chave de artigos publicados	Dados dos projetos: • Autor do projeto • Título • Palavras-chave • Resumo

5.1.2 Tratamento dos dados

São utilizadas técnicas de PLN, focando na utilização dos títulos dos trabalhos (tanto do projeto a ser avaliado como as publicações dos pesquisadores), resumos e palavras-chave. Primeiramente há a remoção de *stop words* definidas a partir de um banco de

dados[1], e a remoção de palavras frequentemente utilizadas em projetos de pesquisa que não trazem grande significado, como por "exemplo", "pesquisa", "análise", "método", "conhecimento", "desempenho".

Em seguida, são removidos acentos e pontuações e, a partir disso, é reduzida ao seu radical e é "colocada" em um saco-de-palavras. Assim, para cada avaliador e cada projeto a ser avaliado criou-se um saco-de-palavras com as principais palavras associados a essa pessoa ou projeto. Outro método utilizado durante o projeto foi a utilização de bigramas, criados antes de serem colocados no saco-de-palavras (Qader W., 2019).

5.2 Criação do grafo

O primeiro procedimento é a criação do grafo onde os avaliadores e projetos avaliados são os vértices, divididos em dois grupos, logo um grafo bipartido. No primeiro momento esse grafo bipartido é completo, ou seja, todos os avaliadores estão ligados a todos os projetos. Cada vértice tem seu saco-de-palavras associado e o valor a ser definido para cada aresta que conecta um avaliador a um projeto será um coeficiente do quão semelhantes são as palavras em cada um, assim temos um grafo bipartido ponderado não direcionado.

5.2.1 Criação dos vértices

No total o grafo tem "número de projetos + número de possíveis avaliadores" vértices. A partir dos dados coletados, como descrito anteriormente, cada vértice tem um atributo com todas as palavras associadas a ele e um segundo atributo indicando se o vértice fará parte da bipartição dos projetos ou avaliadores.

5.2.2 Definição do coeficiente de similaridade

Para definir a relação entre os vértices é utilizado o saco-de-palavras descrito na última Seção 3 onde para cada par de vértices foi estimada a Distância de Jaccard (Niwattanakul S., 2013). Formalmente, para cada par de nós é calculado:

$$intersecção/união, \tag{1}$$

em que *intersecção* refere-se ao número total de termos em comum entre o saco-de-palavras do projeto e do possível avaliador e *união* é o total de termos em ambos sacos-de-palavra. Com isso temos um número entre 0 e 1, que será o peso da aresta entre

[1] A versão utilizada de *stop words* considera 255 palavras, disponível no site https://virtuati.com.br/cliente/knowledgebase/25/Lista-de-stopwords.html.

esse possível avaliador e o projeto. Assim quanto maior a similaridade entre os dois sacos-de-palavras, maior o coeficiente.

5.3 Implementação do emparelhamento máximo

A partir do grafo criado como citado anteriormente, é utilizado um algoritmo de emparelhamento de peso máximo para definir qual avaliador deve avaliar qual projeto, se baseando no coeficiente de semelhança entre avaliador e projeto a ser avaliado. O comando utilizado da biblioteca NetworkX é *nx.max_weight_matching(G)*.

Para fins de simplificação foi realizado o emparelhamento de somente um avaliador por projeto.

5.4 Visualização do grafo

É utilizada a ferramenta Gephi para visualização dos grafos bipartidos gerados. A partir de um comando Python o grafo é exportado em formato *.gexf* após ter sido realizada a remoção de todos os nós de avaliadores isolados do grafo, ou seja, aqueles que não obtiveram nenhuma conexão com os projetos a serem avaliados.

Alguns tratamentos são realizados na ferramenta para a melhor visualização do resultado, entre eles:

- Coloração dos vértices considerando qual sua bipartição.
- Inclusão dos rótulos de cada vértice para identificação tanto de autores de projetos quanto de possíveis avaliadores.
- Coloração e espessura das arestas para refletir o coeficiente de similaridade entre dois vértices - caso a correlação seja alta a aresta será vermelha e mais espessa enquanto as arestas com menor peso são azuis e mais finas.
- Filtro de visualização para mostrar somente as arestas de maior peso.
- Para a distribuição espacial dos vértices são utilizados dois métodos como mostrado na Figura 6: "Event Graph Layout", para que os nós fiquem distribuídos a partir da sua bipartição - autores dos projetos a serem avaliados à esquerda e possíveis avaliadores à direita - e a distribuição "não sobrepor", para que todos os vértices possam ser visualizados.

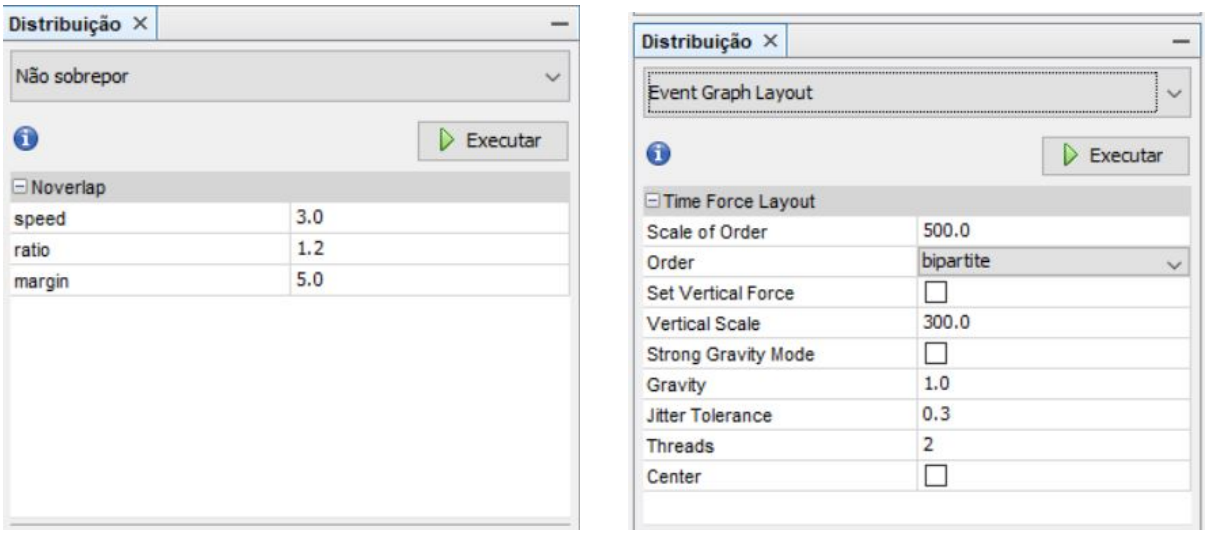

Figura 6: *Distribuições utilizadas para visualização do grafo.*

5.5 Complexidade do código

O código teve sua complexidade analisada nos termos de: número de projetos **p**, número de avaliadores **a**, número total de palavras dos nós dos projetos **pal_p**, número total de palavras dos nós dos avaliadores **pal_a** e o número de palavras a ser retirado (stop words e palavras frequentes) **pal_ret**.

Considerando a complexidade por partes do código:

- Imports e Definições:

$$O(1) \tag{2}$$

- Captura dos dados:

$$O(pal\_p + pal\_a) \tag{3}$$

- Criação banco de palavras a serem retiradas:

$$O(pal\_ret) \tag{4}$$

- Tratamento PLN:

$$O((pal\_p + pal\_a) * pal\_ret) \tag{5}$$

  O tratamento textual mais oneroso é o de retirada das palavras, onde cada palavra do saco-de-palavras de avaliadores e projetos precisam ser comparadas com as palavras a serem retiradas.

- Criação dos nós:

$$O(pal\_p + pal\_a) \tag{6}$$

- Criação das arestas:

$$O(pal\_p * pal\_a) \tag{7}$$

  Por se tratar de um grafo bipartido iremos criar arestas somente entre vértices de diferentes partes. Porém ainda é necessário comparar todas as palavras de cada avaliador com cada projeto para que seja calculado o coeficiente de similaridade entre eles.

- Emparelhamento:

$$O(E * \sqrt{V}) = O(E * \sqrt{p + a}) \tag{8}$$

  Sendo E o número de arestas do grafo. Valor na prática deve ser menor pois antes de realizar o emparelhamento já retiramos do grafo os vértices de possíveis avaliadores que não tiveram correlação com nenhum projeto, logo o número de vértices tende a ser menor que p+a.

- Resultando em um total de:

$$O(pal\_p * pal\_a) \tag{9}$$

  Ou seja, o código é limitado superiormente pelo tempo de criação dos vértices, isso ocorre pois nesse passo é necessário se comparar todas as palavras dos projetos com a dos avaliadores para que se encontre o coeficiente de correlação entre cada dois nós.

6 Resultados

A seguir são apresentados três blocos de resultados, o primeiro consiste em um teste realizado com dados criados para testar o emparelhamento. O segundo mostra diferentes estratégias que foram utilizadas para o tratamento de dados coletados e seus respectivos resultados e comparações. Por fim é discutido um caso de definição de avaliador para projetos de pesquisa da instituição FIOCRUZ/RJ.

6.1 Teste de emparelhamento

Afim de testar o método de emparelhamento além da corretude do código um teste controlado é realizado, com resultados previamente estipulados para validação.

Dados bipartição 1:

- Alberto: maçã
- Beatriz: banana

- Caetano: pera
- Daniela: uva
- Elaine: amora
- Flavio: banana amora

Dados bipartição 2:

- 1: banana ovo amora
- 2: uva leite café
- 3: banana açúcar farinha

Com os dados acima citados temos o seguinte grafo bipartido (Figura 7). Onde é possível observar as correspondências: o nó 2 só tem conexão de similaridade relevante com o nó Daniela, o nó 3 tem sua aresta mais relevante com o nó Beatriz e o nó 1 tem 3 arestas relevantes, sendo que a aresta compartilhada com Flavio é a de maior relevância. Esse resultado condiz com o encontrado nos dados, uma vez que 1 e Flavio tem duas palavras em comum.

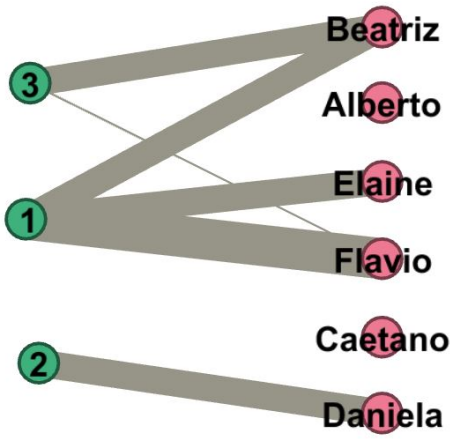

Figura 7: *Grafo resultante do teste de emparelhamento.*

O emparelhamento resultante está apresentado na Figura 8 e mostra que o emparelhamento ocorreu como esperado.

6.2 Emparelhamento com dados coletados

Após seguir os passos indicados no método, obtém-se um grafo bipartido que pode ser visualizado utilizando a ferramente Gephi. Após algumas configurações é possível exibir o

Bipartição 1	Bipartição 2
1	Flavio
2	Daniela
3	Beatriz

Figura 8: *Grafo resultante do teste de emparelhamento.*

grafo de forma em que os nomes dos autores dos projetos estão do lado direito da imagem enquanto os nomes dos avaliadores estão a esquerda. Nas imagens estarão ilustradas somente as arestas mais representativas e somente avaliadores que tiveram algum grau de emparelhamento com algum projeto a ser avaliado.

Para cada um dos projetos a serem avaliados, existem diversas arestas que ligam o projeto aos avaliadores e quanto mais forte o vermelho na figura, maior é o coeficiente de semelhança entre eles.

Além disso, o emparelhamento máximo, realizado a partir de um comando no Python, retorna a lista de projetos com o respectivo avaliador sugerido como será mostrado. O resultado de todos os emparelhamentos está compilado na subseção Resultados Emparelhamento para mais fácil comparação entre todos.

Os testes a seguir tiveram como dados de entrada todos os professores do CMCC da UFABC como avaliadores e oito projetos a serem avaliados. Esses projetos estão divididos entre artigos publicados e projetos de graduação e a coleta dos dados dos projetos foi realizada de forma manual.

6.2.1 Emparelhamento usando dados de entrada sem tratamento de PLN

O primeiro teste realizado não considera nenhuma técnica de PLN, ou seja, os dados origem não são tratados de nenhum modo e todas as palavras foram consideradas. O resultado do grafo encontra-se na Figura 9. É possível observar que existem muitas ligações entre todos os vértices.

Na Figura 10 pode-se observar no detalhe um nó da bipartição dos alunos que submeteram projetos a serem avaliados e suas principais ligações de correspondência com os possíveis avaliadores.

Na figura 1 está o resultado do emparelhamento máximo considerando o peso das arestas pelo coeficiente de similaridade. Com esse resultado pode-se tirar algumas conclusões. Considerando o emparelhamento do nó em destaque na Figura 10 apesar de o nó “Amanda Cruz Francesconi” não ter sido emparelhado com o nó que tem maior correspondência, Marcia Aguiar, a correspondência foi feita com um dos outros possíveis

avaliadores que aparece no grafo, que representam as arestas mais relevantes do grafo.

Uma crítica a esse modelo sem utilizar nenhuma técnica de PLN é em relação aos resultados se observamos os termos em comum, a maior parte dos termos resultantes são palavras sem significado relevante como, por exemplo, pronomes e preposições. Essas palavras não indicam uma similaridade entre o trabalho apresentado e a área de estudo do possível avaliador, portanto podem levar a uma correspondência erronia. Como, por exemplo a correspondência entre os nós "Matheus Porto" e "Itana Stiubiener" em que os únicos termos em comum caem nessa situação: em,uma,de,do,e,a,da. Se esse resultado fosse utilizado o projeto do aluno poderia ser avaliado por uma pesquisadora que não tem conhecimento específico sobre a área de pesquisa do projeto apresentado.

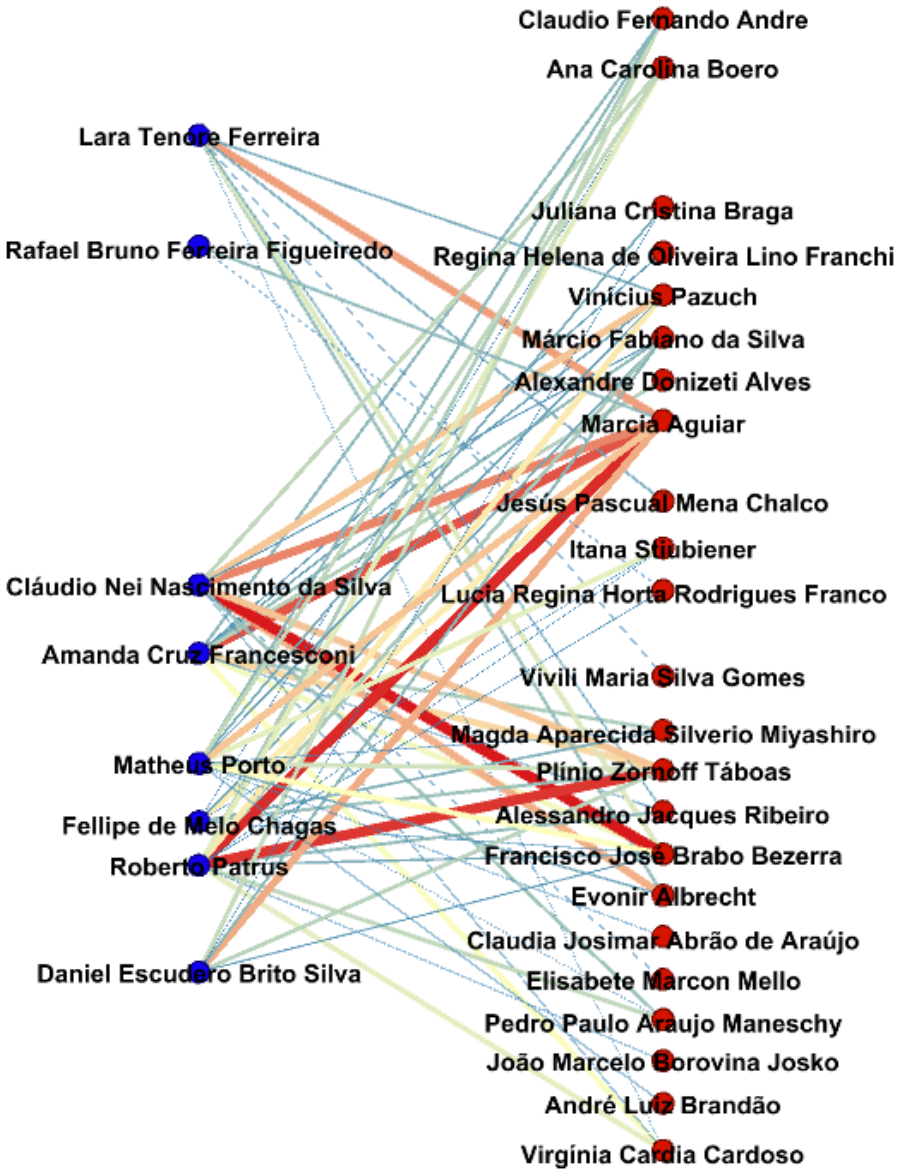

Figura 9: *Grafo bipartido com arestas de maior peso sem tratamento de PLN.*

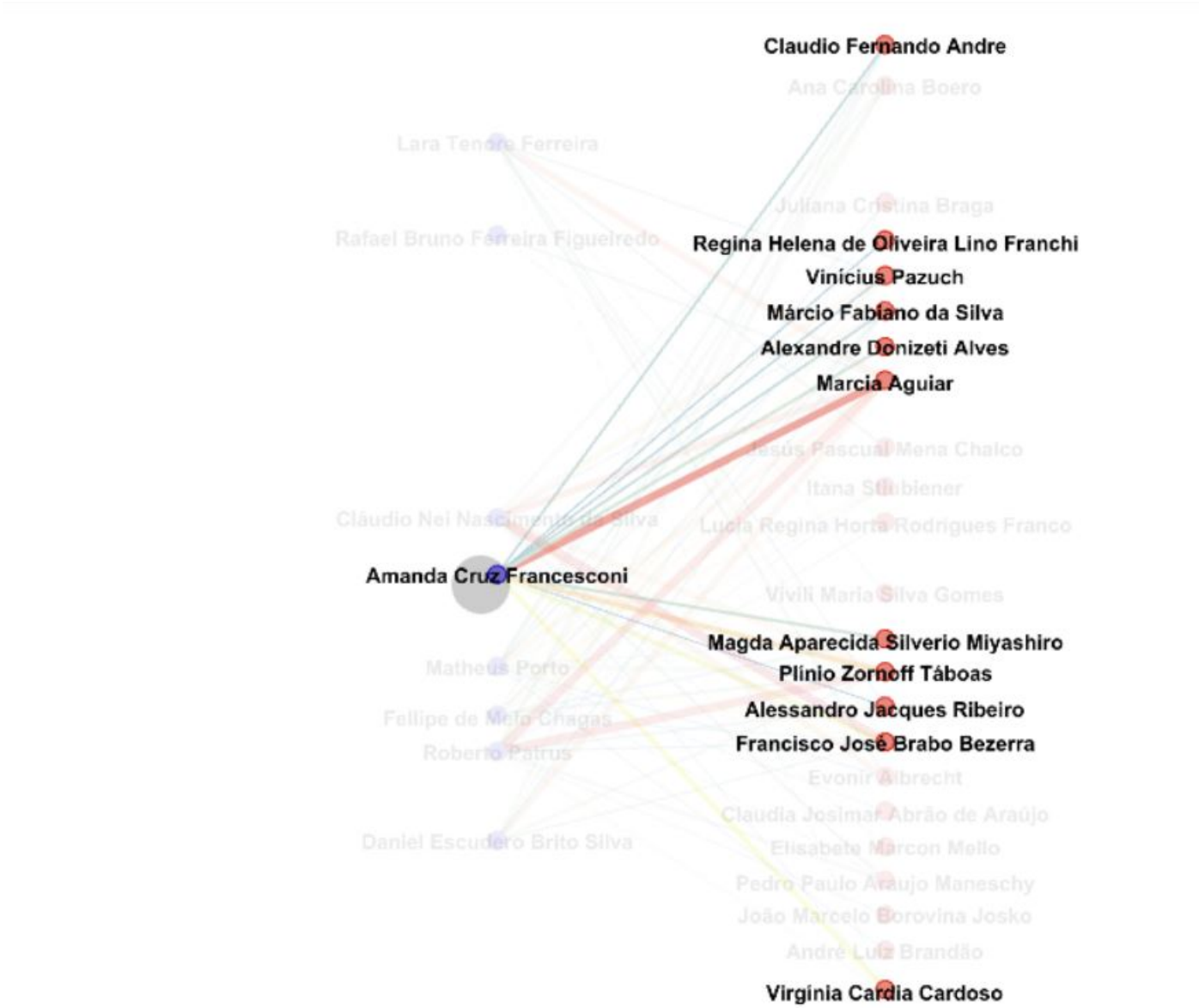

Figura 10: *Destaque do grafo em um aluno que submeteu seu projeto para avaliação - sem tratamento de PLN.*

6.2.2 Emparelhamento usando dados de entrada com tratamento de PLN e bigramas

Para o segundo teste já é considerado todo o tratamento de PLN descrito na seção 5. Além disso, a abordagem considera a criação de bigramas ao invés da técnica do saco-de-palavras utilizada anteriormente. O resultado encontra-se na Figura 11. Nesse caso ao invés de somente comparar as palavras um a um, os termos são comparados dois a dois seguindo a ordem apresentada na origem dos dados. Desse modo o objetivo seria identificar bigramas em comum que trazem mais significado à comparação em relação a palavras soltas como, por exemplo, comparar dois trabalhos com a palavra *inteligência* é menos significativo do que comparar o bigrama *inteligência artificial*, nesse segundo, caso haja correspondência, a correlação entre projeto e avaliador tende a ser mais assertiva.

Como pode-se observar esse grafo apresenta muito menos arestas em relação ao grafo anterior, no teste sem PLN. Isso ocorre pois ao se considerar bigramas a probabilidade de correspondência entre os nós é inferior a quando se compara somente palavras soltas. Outro ponto observado nesse grafo são os nós dos alunos "Matheus Porto" e "Rafael Bruno Ferreira Figueiredo" que nesse grafo, onde só são exibidas as arestas mais relevantes, os dois alunos aparecem sem nenhuma correspondência. Isso ocorre pois esses dois alunos são de áreas diferentes dos professores considerados nesses dados de teste, enquanto os professores são do Centro de Matemática, Computação e Cognição da UFABC esses alunos, e seus projetos, são da área da filosofia e engenharia de instrumentação e robótica respectivamente. Portanto era esperado que esses alunos obtivessem coeficientes de similaridade inferiores aos demais.

Na Figura 12 está o detalhe do grafo focado em apenas um nó de exemplo e reafirmando o ponto citado acima de haver muito menos conexões nessa visão de bigramas, enquanto no teste anterior o nó "Amanda Cruz Francesconi" estava conectado com 11 possíveis avaliadores, nessa visão temos apenas 3 correspondências mais relevantes.

Todos esses pontos citados podem também ser observados na Figura 2, onde estão apresentados os resultados do emparelhamento onde temos coeficientes de correlação (distância de Jaccard), representado pelo peso da aresta, baixíssimos.

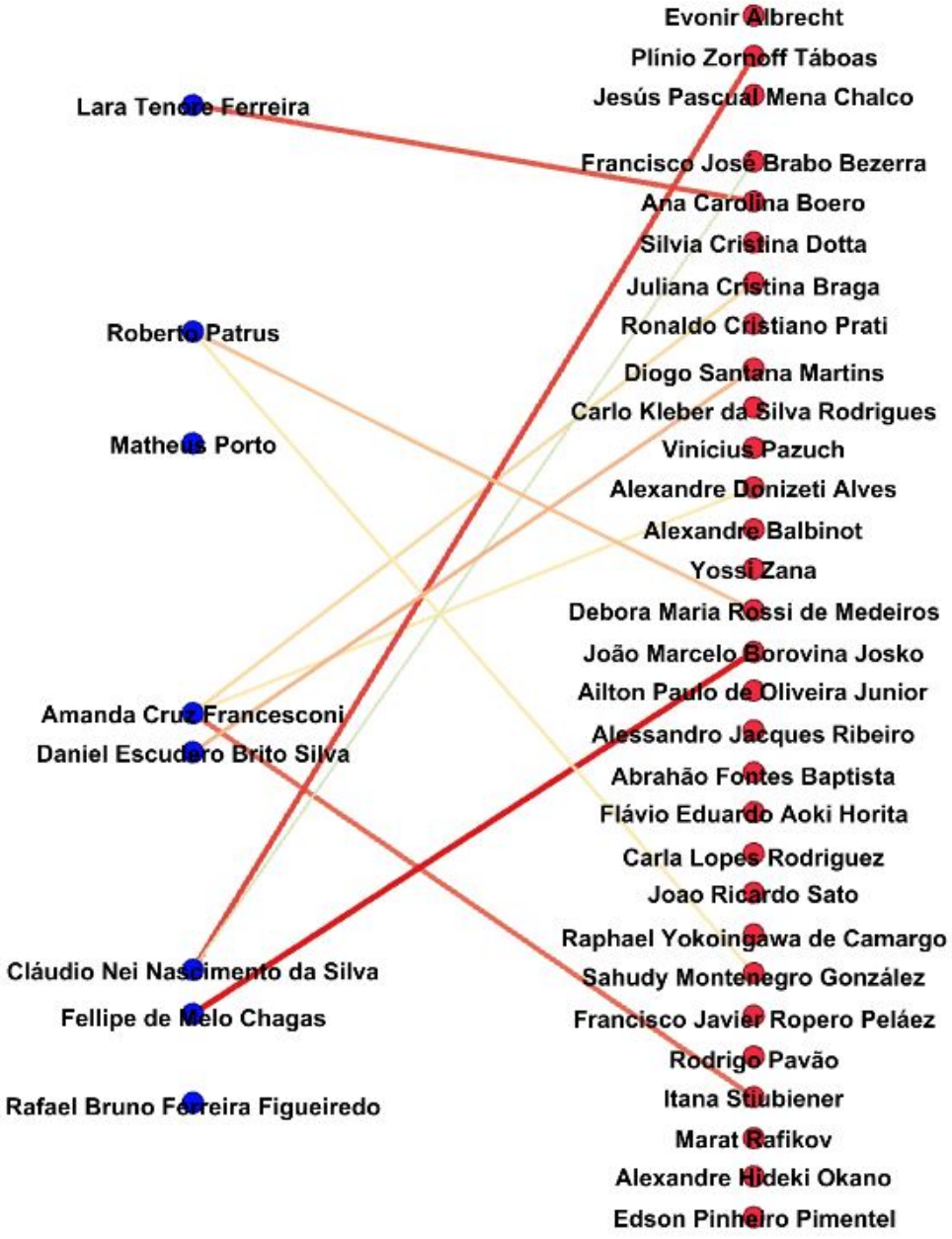

Figura 11: *Grafo bipartido com arestas de maior peso utilizando bigramas.*

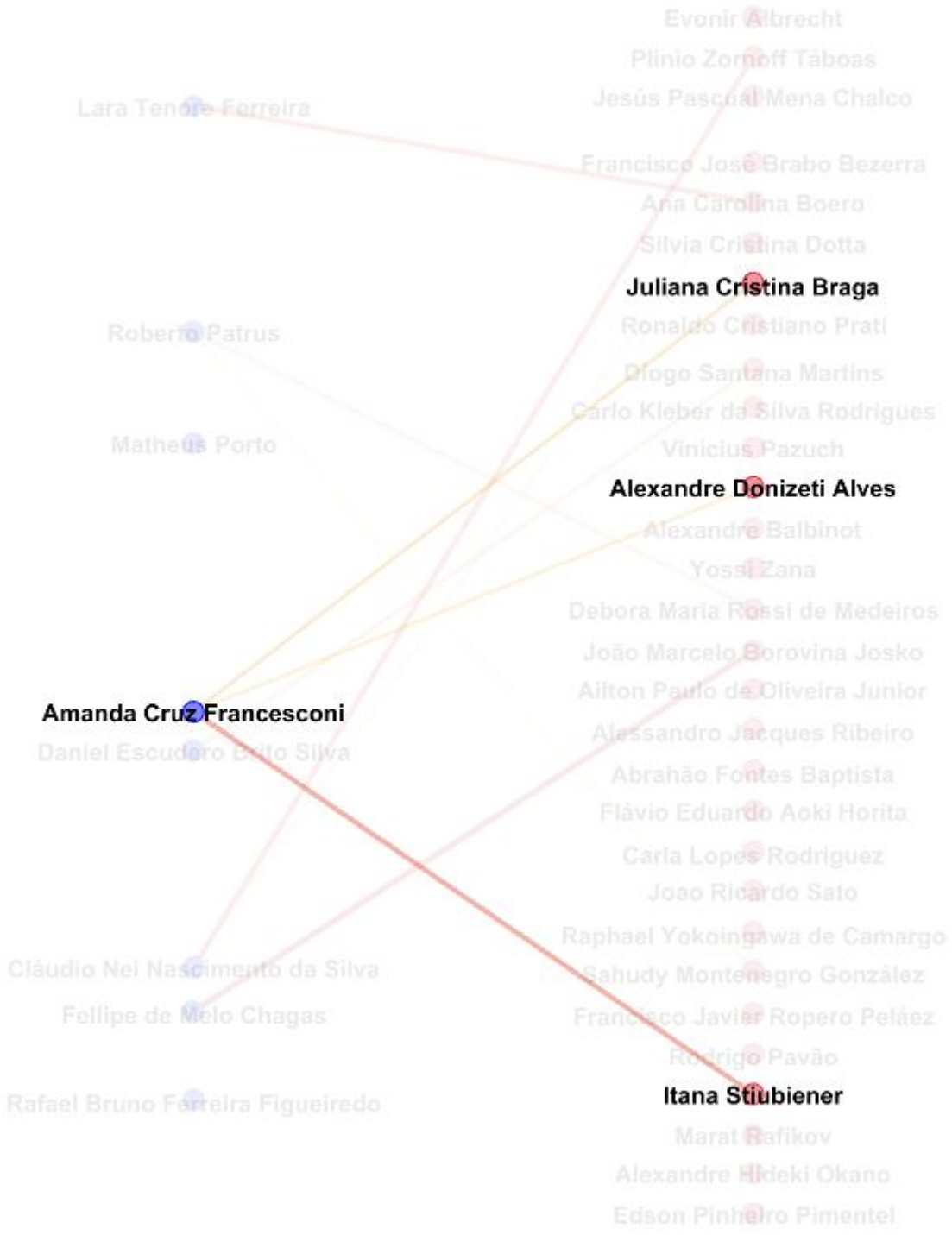

Figura 12: *Destaque do grafo em um aluno que submeteu seu projeto para avaliação - utilizando bigramas.*

6.2.3 Emparelhamento usando dados de entrada com tratamento de PLN sem bigramas

O último emparelhamento utiliza todo o tratamento de PLN e o modelo de saco-de-palavras. O grafo encontra-se na Figura 13. Esse grafo apresenta um número intermediário de arestas em relação aos dois últimos. A observação referente aos nós "Matheus Porto" e "Rafael Bruno Ferreira Figueiredo" feita na última subsecção continua valendo para esse teste, uma vez que esses nós apresentam menos conexões que todos os outros.

Para se chegar a esse resultado diversos testes preliminares foram realizados e algumas melhorias foram: a total retirada de acentos e sinais de pontuações que estavam trazendo ruído para os resultados e um dos tratamentos de PLN foi alterado em relação ao que foi inicialmente proposto. A técnica de retirar 10% das palavras mais frequentes resultou em retirar palavras que tem importante significado e que seriam relevantes ao se fazer a comparação entre os nós. Apesar de ser uma técnica interessante por busca reduzir palavras frequentemente utilizadas em todos os projetos de pesquisa (como por exemplo, análise, algoritmo, fundamento, problema) no caso desse projeto não obtivemos um resultado satisfatório. Afim de obter um resultado similar ao que era esperado com essa técnica foi criado um dicionário de palavras frequentes com 46 termos comuns de serem encontrados nesse âmbito. O dicionário encontra-se no Anexo A.

O destaque para o nó de exemplo encontra-se na Figura 14 onde é possível identificar que existe uma correlação forte com um avaliador específico e na Tabela 3 de resultados vemos que o emparelhamento realmente uniu esses nós que já aparentavam ter uma alta correspondência.

Em suma, o emparelhamento utilizando todos os tratamento de PLN propostos se mostrou o mais interessante para o objetivo almejado.

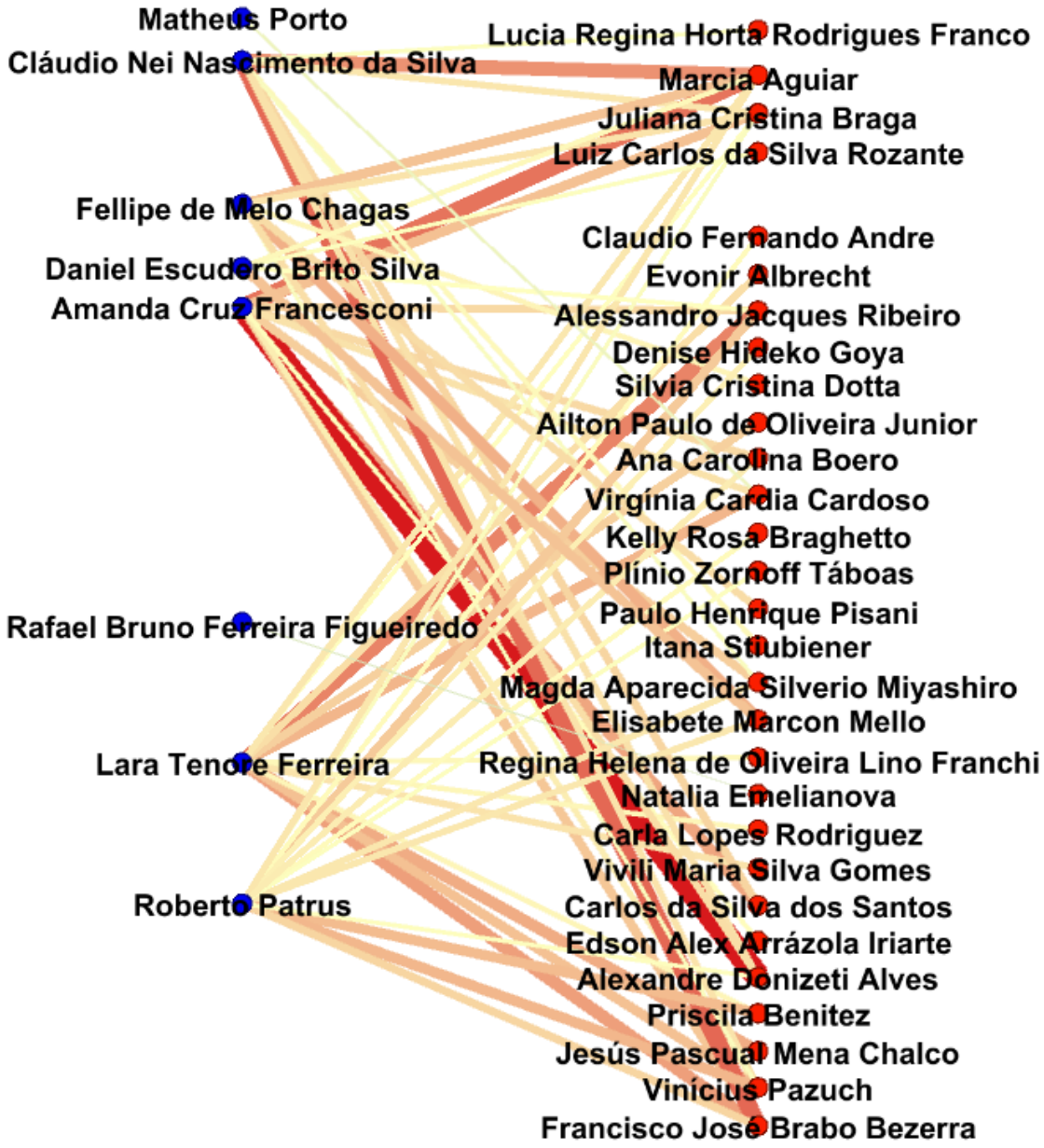

Figura 13: *Grafo bipartido com arestas de maior peso com tratamento de PLN.*

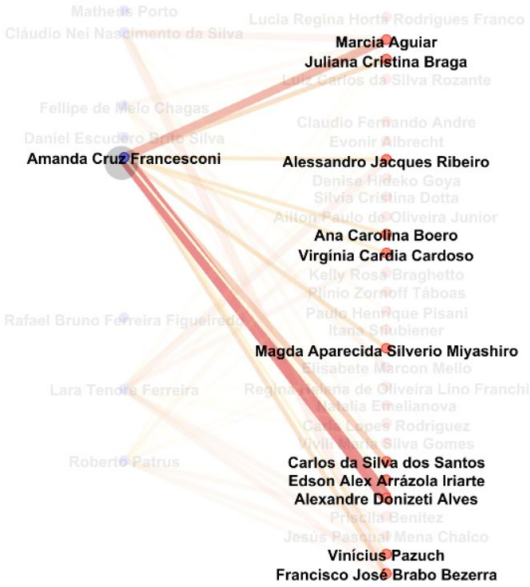

Figura 14: *Destaque do grafo em um aluno que submeteu seu projeto para avaliação - com tratamento de PLN.*

6.2.4 Resultados de emparelhamentos

A seguir são representados os resultados dos emparelhamentos citados na seção anterior, interessante observar que para cada um dos testes a maioria dos resultados do emparelhamento foram distintos (Tabelas 1, 2 e 3).

Tabela 1: *Resultado do emparelhamento projeto - avaliador sem tratamento de PLN.*

Autor projeto	Avaliador sugerido	Peso	Termos em comum
Amanda Cruz Francesconi	Virgínia Cardia Cardoso	0.076	uma, que, vez, a, estudo, em, é, um, para, com, o, no, da, de, projeto, ensino, como, na, processo, e, abordagem, número
Cláudio Nei Nascimento da Silva	Francisco José Brabo Bezerra	0.093	uma, dificuldades, que, A, a, estudo, os, da, partir, em, um, dos, com, o, no, sobre, as, de, onde, na, processo, do, e
Daniel Escudero Brito Silva	Claudio Fernando Andre	0.072	em, de, como, estado, na, para, com, o, computacional, A, O, e, a, da, conhecimento
Fellipe de Melo Chagas	Magda Aparecida Miyashiro	0.067	de, um, utilizando, para, o, a, da
Lara Tenore Ferreira	Marcia Aguiar	0.084	uma, desenvolvido, que, história, desenvolvimento, A, a, em, partir, tarefa, O, para, com, o, no, da, de, como, contexto, na, das, do, e, por
Matheus Porto	Itana Stiubiener	0.075	em, uma, de, do, e, a, da
Rafael Bruno Ferreira Figueiredo	Lucia Regina Rodrigues Franco	0.061	uma, sistema, a, em, controle, O, um, dos, para, com, o, 3, da, as, rede, de, na, problema, do, e, aplicações
Roberto Patrus	Plínio Zornoff Táboas	0.091	de, dos, para, na, História, o, no, e, sobre, da, as,

Tabela 2: *Resultado do emparelhamento projeto - avaliador utilizando bigramas.*

Autor projeto	Avaliador sugerido	Peso	Termos em comum
Amanda Cruz Francesconi	Itana Stiubiener	0.007	estud uma
Cláudio Nei Nascimento da Silva	Plínio Zornoff Táboas	0.008	as estud
Daniel Escudero Brito Silva	Diogo Santana Martins	0.006	arte, estad
Fellipe de Melo Chagas	João Marcelo Borovina Josko	0.008	de
Lara Tenore Ferreira	Ana Carolina Boero	0.007	are uma, pesquis are
Matheus Porto	Evonir Albrecht	0.003	o conceit
Rafael Bruno Ferreira Figueiredo	Joao Ricardo Sato	0.000	conect red
Roberto Patrus	Debora Maria Rossi de Medeiros	0.005	management of

Tabela 3: *Resultado do emparelhamento projeto - avaliador com tratamento de PLN.*

Autor projeto	Avaliador sugerido	Peso	Termos em comum
Amanda Cruz Francesconi	Alexandre Donizeti Alves	0.079	especif, bas, linguag, estud, are, ferrament, natural, process
Cláudio Nei Nascimento da Silva	Francisco José Brabo Bezerra	0.065	exist, conhec, part, estud, signific, onde, compar, dificuldad, process
Daniel Escudero Brito Silva	Elisabete Marcon Mello	0.050	especial, constru, propost, registr
Fellipe de Melo Chagas	Marcia Aguiar	0.050	curricul, bas, pesquis, conhec, process
Lara Tenore Ferreira	Alessandro Jacques Ribeiro	0.061	calcul, compreensa, histor, divers, futur, busc, conjunt, signific, context, projet, impact, pesquis, consider, part, estud, taref, ii, relaco, process
Matheus Porto	Virgínia Cardia Cardoso	0.037	filosof, profund, leitur, critic, construca
Rafael Bruno Ferreira Figueiredo	Natalia Emelianova	0.034	encontr, pont, simpl, font, red, uso, ambient
Roberto Patrus	Vinícius Pazuch	0.052	pesquis, of, histor, review, cientif, uso, teoric, artig, process

6.3 Teste com dados FIOCRUZ/RJ

O método também é utilizado em um teste com dados reais de uma instituição de pesquisa FIOCRUZ no Rio de Janeiro que tem como objetivo fazer a seleção de avaliadores da própria instituição para projetos de pesquisa. Para esse exemplo prático algumas alterações foram necessárias no método. A instituição de pesquisa solicitou que para cada projeto fossem atribuídos três avaliadores porém cada avaliador só teria disponibilidade para receber dois projetos. As resoluções encontradas para essas necessidades são descritas a seguir.

1. **Necessidade de três avaliadores por projeto:** Para se solucionar essa questão foram realizados três emparelhamentos subsequentes onde após cada emparelhamento realizado as arestas já utilizadas eram retiradas do grafo total para que as relações não se repetissem.

2. **Máximo de dois projetos por avaliador:** Considerando que nesse caso foram realizados 3 emparelhamentos, cada professor da instituição só estaria disponível para avaliar 2 projetos, assim foi incluído no grafo um atributo para os nós dos avaliadores onde era acompanhado o número de projetos que já haviam sido designados para ele, assim caso o avaliador já tivesse sido emparelhado nas duas primeiras ele não é considerado para ser emparelhado na terceira rodada.

Os resultados encontram-se na Tabela 4 está representado o resultado de três emparelhamentos realizados com dados reais, onde foi enviado o nome do projeto, qual avaliador sugerido, o coeficiente de similaridade e o número de termos em comum. Os nomes dos projetos e avaliadores foram codificados para preservar a segurança de dados dos participantes.

Tabela 4: *Resultado dos 3 emparelhamentos realizados para um caso real de seleção de avaliadores para projetos de pesquisa*

Candidato	Avaliador 1	Similaridade	Termos em comum	Avaliador 2	Similaridade	Termos em comum	Avaliador 3	Similaridade	Termos em comum
1	A	0.1267	118	E	0.1128	94	A C	0.1116	125
2	B	0.0859	70	A L	0.054	65	Y	0.0988	131
3	C	0.0682	121	A M	0.0997	158	G	0.1236	203
4	D	0.095	62	L	0.0853	49	A B	0.0647	29
5	E	0.0947	76	A	0.0965	88	O	0.0787	143
6	F	0.0702	71	R	0.1239	163	A Q	0.0918	147
7	G	0.1208	186	P	0.0518	69	I	0.0939	131
8	H	0.0947	82	A H	0.1117	79	A H	0.1119	79
9	I	0.0908	114	X	0.0642	76	-	-	-
10	J	0.0516	132	C	0.0595	115	AA	0.0634	95
11	K	0.1155	161	A U	0.1245	135	X	0.0604	79
12	L	0.0851	64	Y	0.1047	136	B	0.0866	69
13	M	0.0176	16	A K	0.0383	47	A L	0.0475	62
14	N	0.1186	155	Q	0.1407	139	A U	0.1534	157
15	O	0.0883	168	T	0.0542	70	A N	0.0722	116
16	P	0.0472	53	A X	0.0273	48	C	0.0537	79
17	Q	0.1309	107	A I	0.1135	93	A E	0.0997	68
18	R	0.1385	148	F	0.0747	57	Z	0.1183	109
19	S	0.1108	81	A B	0.053	33	F	0.0567	42
20	T	0.0594	73	Z	0.1132	112	K	0.1154	148
21	U	0.1491	208	G	0.1272	205	V	0.1056	172
22	V	0.1179	185	A S	0.0838	160	A R	0.0859	162
23	W	0.0887	93	A J	0.1235	153	H	0.0955	98
24	X	0.071	92	A W	0.1117	126	A W	0.1117	126
25	Y	0.0914	129	M	0.0083	7	-	-	-
26	Z	0.1144	123	U	0.1358	174	U	0.1357	174
27	AA	0.0776	84	A C	0.1048	118	N	0.1028	128
28	A B	0.0567	34	A E	0.0929	63	S	0.1068	76
29	A C	0.1117	121	A V	0.108	102	A	0.1055	96
30	A D	0.069	153	A R	0.0827	161	R	0.1302	178
31	A E	0.1064	80	V	0.1183	161	Q	0.1279	114
32	A F	0.1228	136	N	0.1148	141	A F	0.1225	136
33	A G	0.104	103	AA	0.0815	86	A G	0.104	103
34	A H	0.1223	91	S	0.1189	91	L	0.0899	70
35	A I	0.1169	104	A G	0.1058	104	A I	0.118	105
36	A J	0.1397	153	D	0.1039	87	E	0.102	79
37	A K	0.0326	35	A Q	0.0879	115	T	0.0549	61
38	A L	0.0571	61	B	0.0845	58	A O	0.074	56
39	A M	0.0984	151	A N	0.083	144	A P	0.0851	185
40	A N	0.0761	128	O	0.0899	178	J	0.046	106
41	A O	0.0881	76	I	0.1018	119	A M	0.0991	125
42	A P	0.0842	197	J	0.0494	125	-	-	-
43	A Q	0.104	160	A F	0.1132	142	P	0.0568	75
44	A R	0.0995	184	A D	0.0836	177	A D	0.0835	177
45	A S	0.0801	155	A T	0.088	121	A T	0.088	121
46	A T	0.0937	124	A P	0.09	191	A S	0.0811	153
47	A U	0.1419	144	W	0.0914	95	W	0.0914	95
48	A V	0.0993	96	A O	0.0848	78	A V	0.0992	96
49	A W	0.1265	142	K	0.1285	178	A J	0.136	174
50	A X	0.0266	55	H	0.0895	107	D	0.0779	90

7 Trabalhos futuros

Após a conclusão desse projeto foram identificados diversos pontos de melhoria e evolução da ferramenta desenvolvida, a fim de que possa ser utilizada futuramente por instituições de ensino e pesquisa. Os itens estão descritos a seguir:

- Testes com mais técnicas de PLN, como por exemplo outros testes com bigramas que se mostraram promissores porém com baixa correlação.
- Criação de versão iterativa do grafo para auxiliar nas análises de corretude.
- Melhora do código para diminuir complexidade.
- Implementação do algoritmo de emparelhamento a fim de comparação entre resultados.
- Transformar a ferramenta criada em produtiva, com maior facilidade para, a partir das entradas (projetos e nome dos professores), ser possível a rápida visualização da saída (emparelhamento final).
- Inclusão de entradas manuais para calibrar interesse de alguns professores, como, por exemplo, conflitos de interesse.
- Testar diferentes coeficientes de correlação mais robustos

8 Considerações Finais

Com todos os resultados apresentados identifica-se que o melhor emparelhamento gerado é aquele que utiliza todas as técnicas de PLN descritas, remoção de acentos, pontuação, *stop words*, palavras frequentes em pesquisas e projetos, além da aplicação de radicalização em todas as palavras e utilização do método do saco-de-palavras uma vez que a técnica de utilizar bigramas retorna poucas correlações entre trabalhos submetidos para avaliação e currículos de professores.

O resultado obtido no teste com dados reais foi satisfatório e auxiliou na atribuição de avaliadores a projetos em um caso prático. As mudanças realizadas no código para que os resultados se adequassem aos necessários foram benéficas e podem ser incluídas no código principal a fim de se incluir funcionalidades a mais a ferramenta. Uma delas é a de diversos emparelhamentos, para que se designe mais de um professor a cada projeto, pode ser útil em casos onde se é necessário múltiplas avaliações por projeto ou para casos onde o professor indicado não tenha disponibilidade para avaliar o projeto, tendo assim mais uma alternativa para sugestão de emparelhamento.A segunda funcionalidade, que está intrinsecamente relacionada a primeira, pois só com a implementação da primeira a segunda se faz necessária, é a de número máximo de projetos por avaliador, onde se é levado em consideração quanta dedicação será requisitada do professor em avaliar projetos.

Este projeto não considera alguns conceitos humanos como desafeto entre pesquisadores ou se um pesquisador não tem mais interesse em determinada área que já trabalhou anteriormente. Como próximos passos, uma possibilidade seria a inserção de regras manuais que evitem combinações específicas no emparelhamento.

Além do ponto acima, um possível próximo passo seria a criação de um software que realize todo o processo descrito, possibilitando a utilização prática deste.

Espera-se que este projeto auxilie na tomada de decisão dos órgãos responsáveis por fazer a designação de avaliadores a projetos, tornando esse processo menos oneroso e mais assertivo.

Referências

Alves, I. 2021. *Lemmatization vs. stemming: quando usar cada uma?*

Bondy, J. A., & Murty, U.S. 1976. *Graph theory with applications.* The Macmillan Press Ltd.

Checco, A. at al. 2021. Ai-assisted peer review. *Humanities and social sciences communications*, **8**(1), 1–11.

Conix, S, Vaesen K. 2021. Grant writing and grant peer review as questionable research practices. *F1000research*, **10**(1126).

doccumentation, Networkx. $max_weight_matching$.

Edmonds, J. 1965a. Matching and a polyhedron with 0,l vertices. *Pages 125–130 of: J. res. n.*, vol. 69.

Edmonds, J. 1965b. Path, trees and flowers. *Pages 449–467 of: Can. j. math*, vol. 17.

Eisenstein, J. 2019. *Introduction to natural language processing.* MIT press.

Forsberg, E. 2022. Peer review in academia. *Pages 3–36 of: Peer review in an era of evaluation.* Palgrave Macmillan, Cham.

Galil, Z. 1986. Efficient algorithms for finding maximum matching in graphs. *Pages 23–38 of: Computing surveys*, vol. 18.

Hachaj T., Ogiela, M. 2018. What can be learned from bigrams analysis of messages in social network? *Pages 1–4 of: 2018 11th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei).*

Jurafsky, D., & Martin, J. 2000. *Speech and language processing.* Prentice Hall.

Lane, H. 2019. *Natural language processing in action: Understanding, analyzing, and generating text with python.* Manning.

Mena-Chalco, J. et al. 2012. Minerando e caracterizando dados de currículos lattes. *Pages 1–12 of: Brazilian workshop on social network analysis and mining - brasnam.*

Mrowinski, MJ, et al. 2017. Artificial intelligence in peer review: How can evolutionary computation support journal editors? *Plos one.*

Niwattanakul S., at al. 2013. Using of jaccard coefficient for keywords similarity. *Pages 380–384 of: Proceedings of the international multiconference of engineers and computer scientists*, vol. 1.

Orengo, V., Huyck C. 2001. A stemming algorithm for the portuguese language. *Pages 186–193 of: Proceedings eighth symposium on string processing and information retrieval*.

Patrus R., Dantas D., Shigaki H. 2016. Pesquisar é preciso. publicar não é preciso: história e controvérsias sobre a avaliação por pares. *Avaliação*, **21**(3).

Prestes, E. 2020. *Introdução à teoria dos grafos*.

Qader W., at al. 2019. An overview of bag of words;importance, implementation, applications, and challenges. *Pages 200–204 of: 2019 international engineering conference (iec)*.

Rajaraman, A. 2014. Mining of massive datasets. *Page 476 of: Cambridge university press*.

Serra F., Ferreira M., et al. 2007. Publicar é difícil ou faltam competências? o desafio de pesquisar e publicar em revistas científicas na visão de editores e revisores internacionais. *Encontro de ensino e pesquisa em administração e contabilidade*, **9**(4).

Silva C., Machado S. 2016. A revisão por pares a partir da percepção dos editores: um estudo comparativo em revistas brasileiras, espanholas e mexicanas. *Rdbci: Revista digital de biblioteconomia e ciência da informação*, **14**(1), 126–143.

Terán, C. 2011. Aspectos éticos de las comunicaciones científicas. *Galícia clínica*, **72**(4), 169–179.

Zhang, Guangyao, Xu, Shenmeng, Sun, Yao, Jiang, Chunlin, & Wang, Xianwen. 2022. Understanding the peer review endeavor in scientific publishing. *Journal of informetrics*, **16**(2), 101264.

A Dicionário de palavras frequentes em pesquisa

A seguir são apresentados os 46 termos que foram retirados por serem termos comuns e não terem significativa importância no contexto de entender qual o assunto da pesquisa ou artigo. As palavras foram radicalizadas para que todas as variações fossem retiradas. Como por exemplo o termo *analis* pode representar as palavras: analise, analises, analisando, analisado.

As palavras foram escolhidas a partir de uma análise de todo conjunto de palavras de entrada, onde foram classificadas as 10% palavras mais frequentes. A partir dessas foram escolhidas as palavras que seriam retiradas.

1. algoritm
2. alternativ
3. amostr
4. analis
5. avaliaca
6. caracterist
7. caracterizaca
8. cienc
9. classificaca
10. colaborativ
11. conhecim
12. desempenh
13. desenvolv
14. disciplin
15. ensin
16. ensino
17. equaca
18. estrutur
19. experimental
20. fundament
21. fundamental
22. implantaca
23. importanc
24. influenc
25. informaca
26. linguagens
27. metod
28. metodolog
29. model
30. objet
31. paramet
32. performanc
33. period
34. perspectiv
35. principi
36. problem
37. produca
38. relaca
39. revisa
40. sistem
41. soluca
42. tecnolog
43. topic
44. trabalh
45. utiliz
46. utilizaca