Pós-graduação no Brasil de 1987 a 2021: uma Análise dos Dados

Francisco Neves; Gabriel Fialho

July 26, 2025

1 Introdução

A pós-graduação no Brasil desempenha um papel crucial no desenvolvimento científico e tecnológico do país. Nas últimas décadas, houve um aumento significativo no investimento em pesquisa e educação de pós-graduação, resultando em um crescimento expressivo no número de trabalhos de dissertações e teses defendidos. Isso pode ser observado na Figure 1.

O objetivo deste relatório é analisar o comportamento e compreender aspectos qualitativos da pós-graduação no Brasil, particularmente o relacionamento entre as *knowledge areas*. Para isso, utilizaremos técnicas de mineração de dados, como regras de associação, redes neurais para classificação e cálculo de distância entre as representações dos trabalhos para agrupamento. Essas técnicas nos permitirão extrair padrões e insights relevantes dos dados.

Para realizar essa análise, utilizaremos dados públicos fornecidos pelo Ministério da Educação (MEC) [1]. Ao identificar padrões e relações entre as *knowledge areas*, poderemos ter uma compreensão mais aprofundada da estrutura e da interdisciplinaridade da produção científica brasileira na pós-graduação.

O banco de dados utilizado neste estudo abrange de 1987 a 2021, tendo, após a limpeza dos dados, 1.240.034 entradas, representando as dissertações e teses de pós-graduação *stricto-sensu* defendidas nesse período.

Este relatório está estruturado da seguinte maneira:

- **Seção 1**: Introdução;
- **Seção 2**: Apresentação dos dados, incluindo atributos, características, limitações e técnicas de limpeza;
- **Seção 3**: Classificação e agrupamento dos trabalhos;
- **Seção 4**: Mineração de padrões nas keywords;
- **Seção 5**: Conclusões.

[1] MEC: http://www.mec.gov.br

2 Dados

Os dados utilizados neste projeto foram obtidos a partir dos dados abertos do Ministério da Educação (MEC). No entanto, nem todas as features disponíveis no conjunto de dados original são relevantes para este trabalho. Portanto, serão consideradas apenas as seguintes features:

2.1 defense date

A feature "defense date" se refere à data em que os trabalhos de pós-graduação foram defendidos. Essa informação abrange o período de 1987 a 2021 e foi formatada de acordo com um padrão específico para facilitar a análise. Foram excluídas as datas anteriores a 1987 e posteriores a 2021, a fim de garantir a consistência dos dados. A distribuição do número de trabalhos por ano mostra um crescimento aproximadamente exponencial ao longo do tempo, como ilustrado na Figure 1.

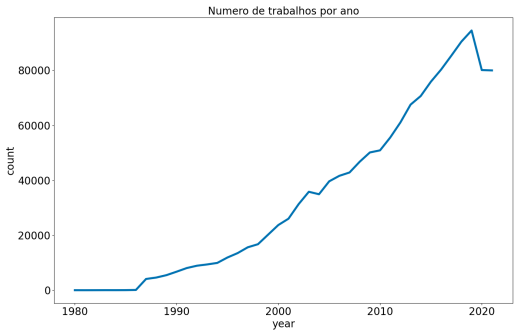

Figure 1: Numero de trabalhos por ano

2.2 knowledge area

A feature "knowledge area" descreve a àrea de conhecimento à qual cada trabalho de pós-graduação pertence. Durante a limpeza dos dados, foram removidas áreas com menos de 100 trabalhos, consideradas irrelevantes para as tarefas de mineração de dados deste projeto. Após essa etapa, restaram 175 valores distintos para essa feature.

No entanto, é importante destacar que há um desequilíbrio significativo entre as knowledge areas, como mostra a distribuição das áreas na Figure 2. A maioria das áreas possui um número relativamente baixo de trabalhos em comparação com outras áreas. Esse desequilíbrio deve ser considerado durante a análise dos dados, a fim de evitar distorções que favoreçam áreas com maior número de trabalhos.

Embora haja um desequilíbrio, ele não segue uma lei de potência (power law), como evidenciado na Figura 4. Portanto, é necessário abordar esse desequilíbrio de forma adequada para garantir que a análise seja representativa de todas as áreas.

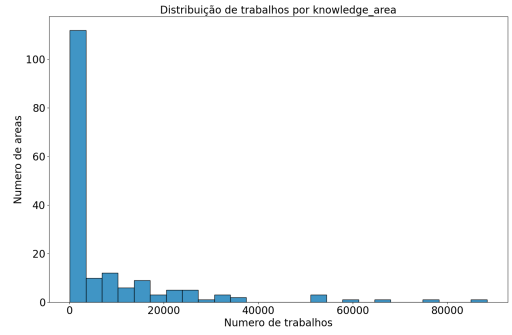

Figure 2: Histograma do numero de trabalhos por knowledge area

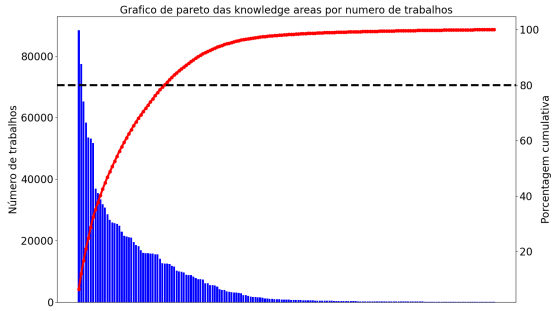

Figure 3: Pareto do numero de trabalhos por knowledge area

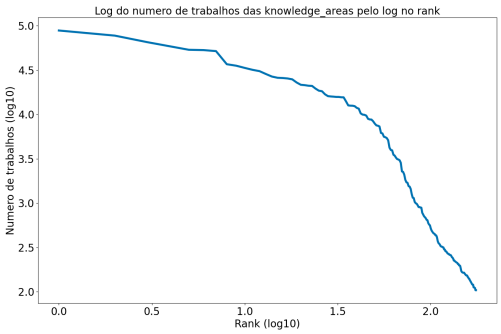

Figure 4: Log/Log das knowledge areas

2.3 keywords

No atributo "keywords", estão armazenadas as palavras-chave selecionadas para descrever cada trabalho. Essas keywords passaram por várias etapas de pré-processamento para garantir a consistência e a qualidade dos dados. As etapas de limpeza aplicadas foram as seguintes:

- **Remoção de símbolos não alfanuméricos**: Símbolos não alfanuméricos, como pontuação e caracteres especiais, foram removidos das keywords para manter apenas os caracteres relevantes.

- **Remoção de stopwords em português e inglês**: Foram removidas as stopwords, ou seja, palavras comuns em português e inglês que não possuem um significado distintivo. Essas palavras não contribuem significativamente para a análise de texto.

- **Stemming**: Foi aplicada a técnica de stemming para reduzir as keywords à sua forma raiz, removendo sufixos e considerando variações da mesma palavra como uma forma única. Isso ajuda a reduzir a quantidade de keywords e agrupar termos similares.

- **Remoção de frequência 1**: keywords que ocorrem apenas uma vez foram removidas, pois provavelmente são erros de digitação ou têm pouca relevância para a análise.

- **Remoção por entropia**: As keywords foram avaliadas com base em sua entropia em relação às knowledge areas. A entropia mede a aleatoriedade de uma distribuição e, nesse caso, indica o quão igualmente as keywords são usadas em diferentes áreas. keywords com alta entropia são usadas em várias áreas, sendo consideradas menos informativas. Foi definido um limite de 4.5 para a entropia a fim de remover keywords com pouca contribuição. A Figure 5 indica o histograma das entropias

- **Remoção de keywords redundantes**: keywords redundantes foram identificadas com base em suas probabilidades condicionais. Se duas keywords ocorrem frequentemente juntas, uma delas é considerada redundante e removida. Utilizou-se um limiar de 0.9 para considerar que um palavra é dedante.

- **Remoção de frequência 1 novamente**: Por fim, repetiu-se a remoção de keywords com frequência 1.

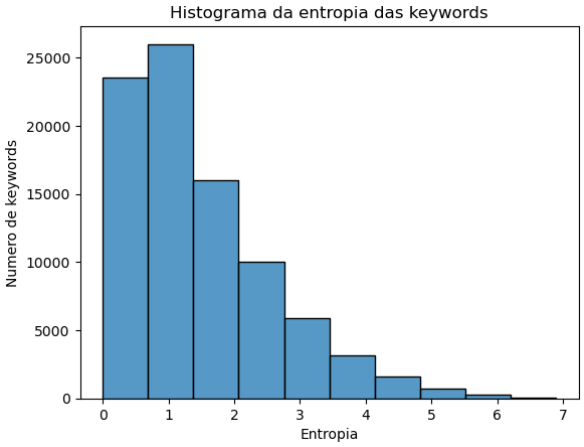

Figure 5: Entropia das keywords

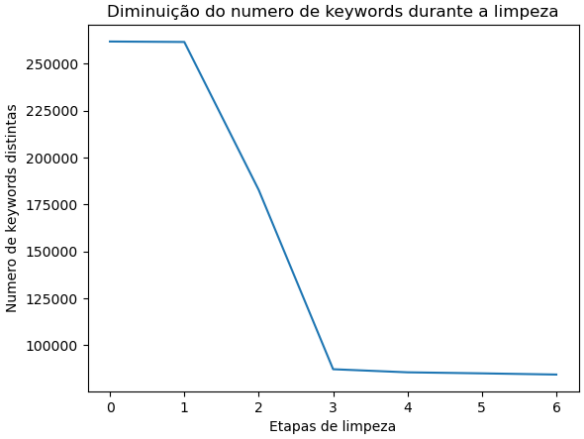

Figure 6: Processo de limpeza das keywords

Durante o pré-processamento dos dados, observou-se que as etapas que mais eliminaram keywords foram o stemming e a primeira remoção de frequência 1.

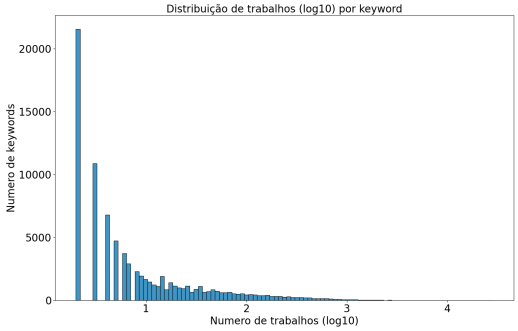

Figure 7: Histograma do numero de trabalhos por keyword

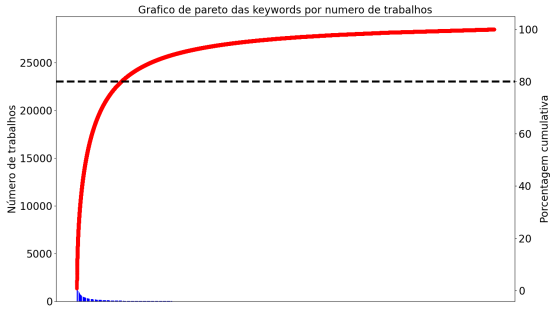

Figure 8: Pareto do numero de trabalhos por keyword

Ao analisar a distribuição dos trabalhos por keywords, como mostrado na Figura 7, e a representação da distribuição de Pareto, ilustrada na Figura 8, é possível notar um padrão distinto em comparação com a distribuição por knowledge areas. A distribuição de trabalhos por keywords possui uma cauda mais pesada, indicando que algumas keywords estão associadas a um grande número de trabalhos, enquanto muitas keywords estão relacionadas a um número reduzido de trabalhos.

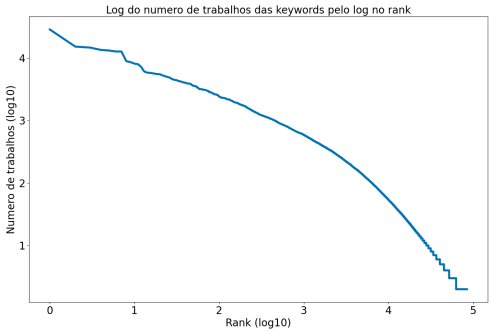

Figure 9: Log/Log das keywords

Ao analisar a Figura 9, verifica-se que a distribuição de keywords não segue um comportamento de lei de potência (power law), embora apresente maior semelhança com esse tipo de distribuição em comparação com a distribuição por knowledge areas.

Apesar dessa distribuição desigual das keywords, isso não deve ser um problema significativo nas etapas de classificação e agrupamento, especialmente considerando a entropia das keywords. No caso do agrupamento, as keywords não são utilizadas diretamente, tornando sua distribuição menos relevante para essa etapa.

No entanto, para a mineração de itemsets, essa distribuição desigual das keywords pode representar um desafio, pois a métrica principal é o suporte (support), e algumas keywords possuem um suporte desproporcional em relação a outras.

3 Classificação e Agrupamento

Os dados textuais apresentam o desafio da alta dimensionalidade devido à grande quantidade de palavras. Uma solução comum é usar representações compactas aprendidas durante o treinamento do modelo, como o word-to-vect.

Neste trabalho, optamos por usar uma rede neural para classificar as keywords em knowledge areas. Uma rede neural profunda é uma arquitetura de aprendizado de máquina composta por várias camadas sequenciais que aplicam multiplicação de matrizes e funções de ativação não lineares. Essas camadas transformam a entrada de maneira não linear, ajustando os pesos para minimizar a diferença entre a transformação da entrada e a saída desejada. Cada camada produz uma representação da entrada que depende da saída desejada e da convergência da rede durante o treinamento.

Em resumo, neste estudo, usaremos as keywords como entrada para uma rede neural que realizará a classificação dos trabalhos, e as representações aprendidas serão usadas para o agrupamento.

3.1 Classificação

A estrutura geral e o fluxo de dados da rede neural utilizada estão representados na Figura 10.

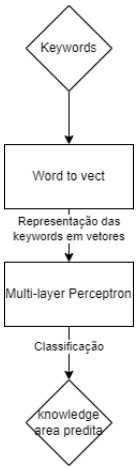

Figure 10: Fluxograma da rede neural

A primeira etapa do processo consistiu em converter as keywords em vetores usando a técnica word-to-vec. Isso permitiu a representação numérica das palavras, capturando suas relações semânticas e contextuais.

Em seguida, a representação do trabalho foi obtida calculando a média dos vetores correspondentes às keywords presentes nele. Essa média forneceu uma representação agregada e compacta das keywords, capturando suas principais características e tópicos abordados.

Para a classificação da knowledge area, utilizou-se um perceptron multicamadas (MLP). O MLP é composto por uma ou mais camadas ocultas, permitindo a aprendizagem de características complexas e não lineares. A camada de saída do MLP foi configurada com o mesmo número de neurônios correspondente ao número de knowledge areas existentes.

A rede neural foi treinada usando a função de perda CrossEntropyLoss, que mede a discrepância entre as predições da rede e as classes reais das knowledge areas. Também foram aplicados pesos inversamente proporcionais ao número de trabalhos em cada knowledge area, para equilibrar a importância das áreas menos representadas.

As estatísticas apresentadas nas tabelas referem-se ao conjunto de teste usado para avaliar o desempenho dos modelos. A Tabela 1 mostra as métricas de precisão, revocação e F1-score para a rede neural e outros dois modelos comparativos. As métricas foram calculadas usando uma abordagem macro, que calcula as médias para cada knowledge area, fornecendo uma visão geral do desempenho em todas as áreas.

Modelo	**Precisão**	**Revocação**	**F1-score**
Random Forest	0.18	0.12	0.12
SVM	0.18	0.14	0.14
Rede Neural	0.19	0.14	0.16

Table 1: Metricas dos modelos

Os modelos foram treinados na partição de treinamento dos dados e avaliados na partição de teste. Para a tarefa de classificação, as representações dos trabalhos obtidas pelo word-to-vec foram usadas como entrada para os modelos.

Ao analisar as métricas, observamos que os modelos tiveram desempenho bastante próximo. A rede neural se destacou ligeiramente em relação aos outros dois modelos. No entanto, é importante notar que a rede neural também aprendeu as representações dos trabalhos usando o word-to-vec, o que adicionou complexidade adicional ao seu treinamento.

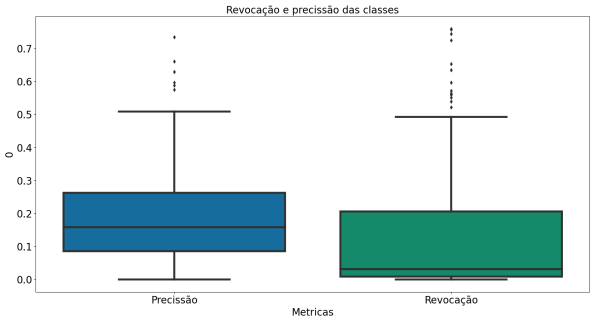

Figure 11: Box plot da precissão e da revocação

Ao examinar a performance da rede neural, podemos observar algumas observações importantes. A precisão do modelo não é influenciada pelo número de trabalhos em cada knowledge area, como pode ser visto na Figura 12.

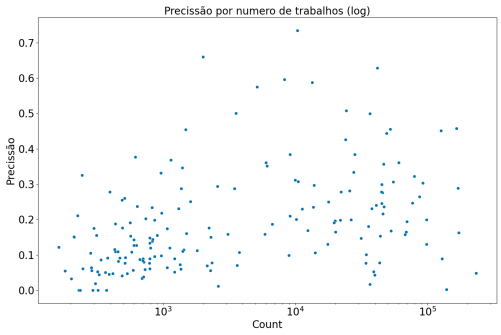

Figure 12: Precisão por numero de trabalhos (log)

No entanto, a revocação do modelo mostrou uma relação direta com o número de trabalhos em cada knowledge area. Quanto maior o número de trabalhos, maior tende a ser a revocação do modelo para essa área específica, como ilustrado na Figura 13.

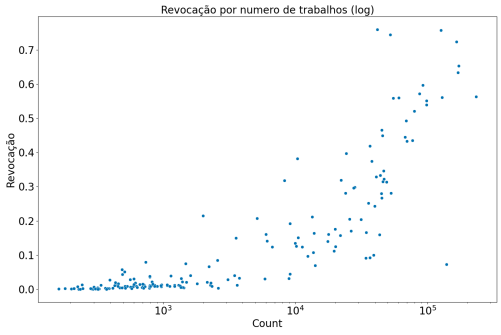

Figure 13: Revocação por numero de trabalhos (log)

Observamos que o modelo pareceu focar na precisão, apresentando uma distribuição relativamente equilibrada entre as classes. No entanto, a revocação do modelo foi afetada negativamente. Esse fenômeno pode ser atribuído a uma estratégia de compensação adotada pelo modelo para lidar com classes infrequentes. Em alguns casos, para aumentar a revocação em classes de baixa representatividade, o modelo pode classificar incorretamente trabalhos em classes mais frequentes. Isso pode ser observado nos mapas de calor das Figuras 14 e 15, onde são destacadas classificações errôneas em áreas mais populares.

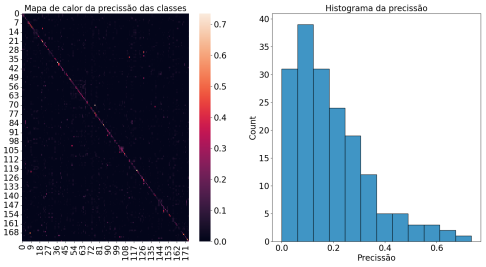

Figure 14: Mapa de calor e histograma da precisão

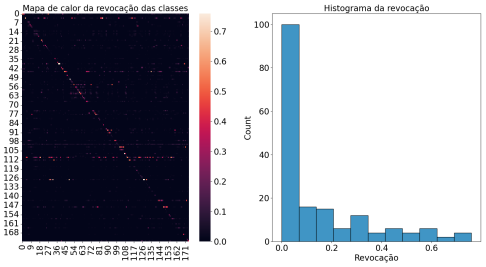

Figure 15: Mapa de calor e histograma de revocação

Além disso, ao analisar os mapas de calor, podemos identificar a presença de "blocos" distintos, indicando a existência de grupos de knowledge areas que o modelo tem bastante dificuldade de distinguir.

3.2 Agrupamento

O processo de agrupamento envolveu a utilização do algoritmo K-means com a representação dos trabalhos obtida a partir das keywords por meio da rede neural/word-to-vec. O objetivo era identificar padrões nas knowledge areas.

Para representar as knowledge areas, utilizamos um vetor médio calculado a partir da média dos trabalhos em cada área. Optamos por ter 11 clusters com base na maior medida de silhueta obtida ao agrupar os vetores médios, como mostrado na Figura 16.

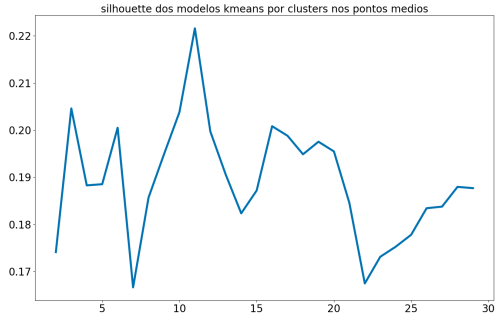

Figure 16: Silhoutte do modelo por clusters nos pontos medios

Na visualização do agrupamento por vetor médio, plotamos os pontos e grupos em duas dimensões usando a técnica de redução de dimensionalidade PCA. Observamos uma distribuição semelhante de trabalhos em cada grupo, conforme apresentado na Figura 18. Além disso, o número de knowledge areas por grupo também foi similar, como ilustrado na Figura 19.

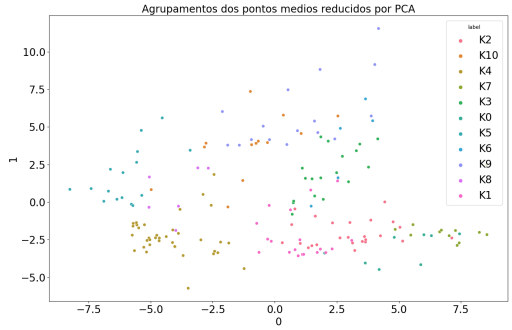

Figure 17: Distribuição de trabalhos por knowledge area

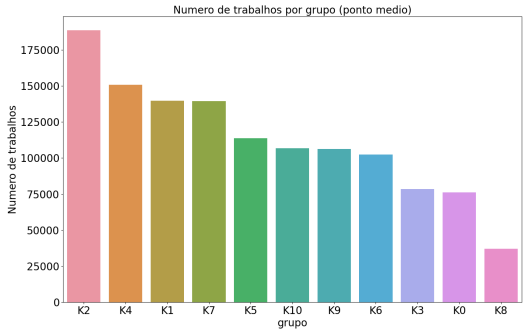

Figure 18: Numero de trabalho por grupo

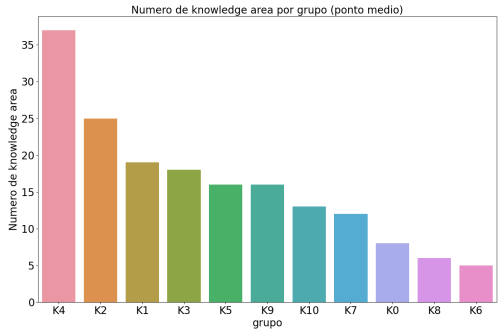

Figure 19: Numero de knowledge area por grupo

A seguir, são apresentados exemplos de knowledge areas em cada grupo:

- K2 = administracao; administracao de empresas; administracao publica; arqueologia; arquitetura e urbanismo
- K10 = agronomia; botanica; ecologia; fitopatologia; fitossanidade
- K4 = anatomia; anatomia patologica e patologia clinica; biologia molecular; cancerologia; cardiologia
- K7 = antropologia; artes; comunicacao; historia; letras
- K3 = astronomia; ciencia do solo; ciencias ambientais; engenharia agricola; engenharia civil
- K0 = biblioteconomia; ciencia da informacao; ciencia politica; direito; direito publico
- K5 = biofisica; biologia geral; bioquimica; biotecnologia; clinica veterinaria
- K6 = ciencia da computacao; engenharia aeroespacial; engenharia de producao; engenharia eletrica; engenharia mecanica
- K9 = ciencia e tecnologia de alimentos; engenharia de alimentos; engenharia de materiais e metalurgica; engenharia nuclear; engenharia quimica
- K8 = cirurgia buco-maxilo-facial; clinica odontologica; endodontia; odontologia; odontopediatria
- K1 = educacao; educacao de adultos; educacao especial; ensino; ensino de ciencias e matematica

Os grupos resultantes reuniram knowledge areas semelhantes. Por exemplo, o grupo K8 engloba áreas relacionadas ao estudo de plantas, meio ambiente e interações entre seres vivos, enquanto o grupo K3 abrange áreas ligadas ao estudo do corpo humano, suas estruturas, doenças e tratamentos.

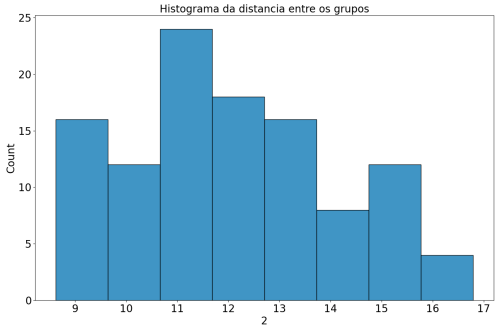

Figure 20: Histograma da distancia média entre os grupos

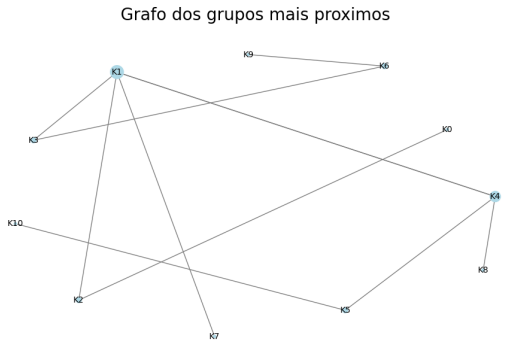

Figure 21: Histograma da distancia média entre os grupos

Calculamos a distância entre os grupos com base na média das distâncias entre os vetores médios de cada grupo. A distribuição das distâncias pode ser observada na Figura 20. O grafo dos grupos, mostrado na Figura 21, ilustra as relações de proximidade entre eles. As arestas do grafo representam a proximidade entre os grupos, indicando quais estão mais próximos uns dos outros.

Destaca-se que o grupo K1 atua como um "hub" no contexto do agrupamento, estando conectado a vários outros grupos. Esse grupo está associado a áreas relacionadas à educação, o que pode explicar sua proximidade com diferentes grupos.

Também é relevante a proximidade entre os grupos K9 e K6, que representam áreas das ciências exatas e estão próximos à engenharia. Isso indica uma possível sobreposição temática ou interdisciplinaridade entre esses grupos, sugerindo que existem temas comuns entre as áreas de estudo relacionadas à engenharia e às ciências exatas..

4 Mineração de padrões nas keywords

O objetivo dessa seção é extrair padrões nos dados através de keywords e knowledge areas presentes nos trabalhos.

Nesta parte, foi decidido trabalhar apenas com as 30 knowledge areas que apresentaram o maior número de trabalhos. Os motivos para essa escolha foram focar nas áreas mais relevantes e lidar com um volume menor de dados. Entretanto, é importante destacar que as áreas com menos trabalhos não estão representadas nesse etapa.

O conjunto de dados original possui 1366973 transações, representando trabalhos de pós graduação. Após filtrar as transações das 30 knowledge areas com mais transações, o novo conjunto de dados obtido contém 798227 transações, ou seja, cerca de 10% das knowledge areas concentram aproximadamente 60% dos trabalhos de pós graduação.

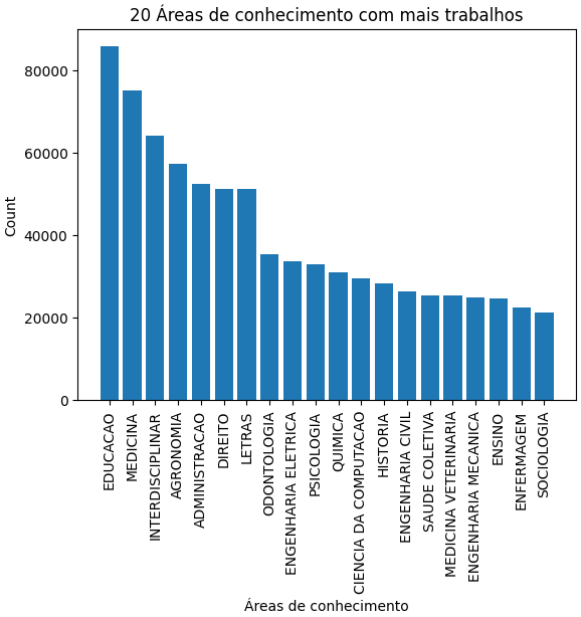

Figure 22: Distribuição de trabalhos por knowledge area

4.1 Visualizações

1. **Matriz de Similaridade entre knowledge areas**:

   Para avaliar a similaridade entre as knowledge areas através de keywords, foram selecionadas as 100 principais keywords de cada área. Com base nessas keywords, uma matriz de similaridade foi criada usando o índice de Jaccard. Esse índice mede a similaridade entre duas áreas de acordo com a quantidade de keywords em comum entre as 100 principais keywords de cada área.

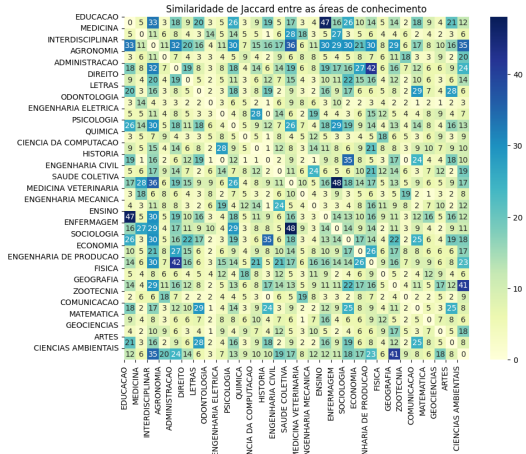

Figure 23: Matriz de similaridade de Jaccard entre knowledge areas

2. **Histograma das Combinações de knowledge areas mais Similares**:

Após a criação da matriz de similaridade, um histograma foi gerado para exibir as 20 combinações de knowledge areas com maior similaridade. Nesse histograma, observou-se que a área 'INTERDISCIPLINAR' ocorre em várias combinações com boa similaridade. Isso pode ser explicado pelo fato de 'INTERDISCIPLINAR' ser uma área que abrange várias disciplinas diferentes.

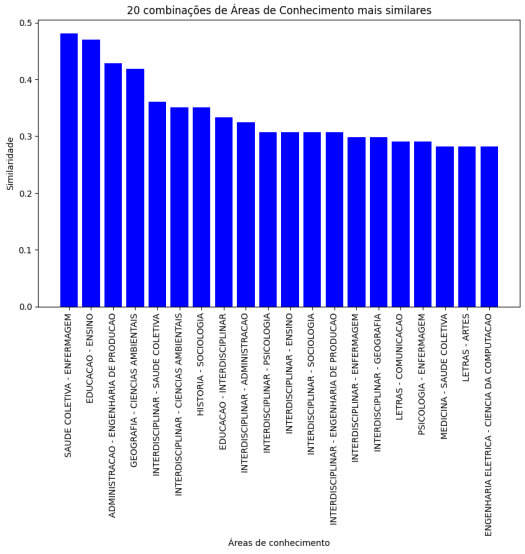

Figure 24: Combinações de knowledge areas mais similares

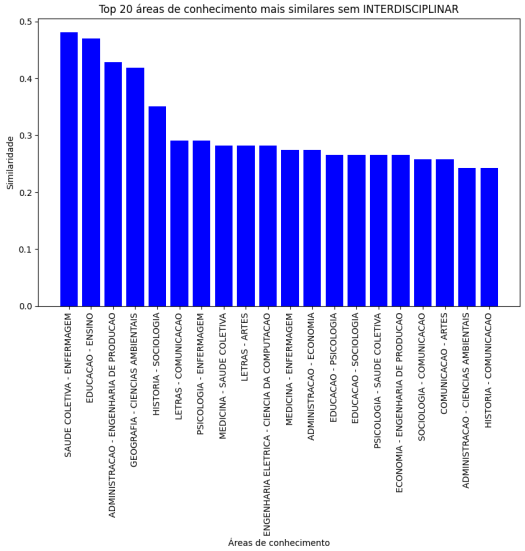

Figure 25: Combinações de knowledge areas mais similares desconsiderando 'INTERDISCIPLINAR'

3. **Grafo de similaridades**:

   A partir da matriz de similaridades obtida, foi construído um grafo no qual os nós são as knowledge areas, o tamanho do nó é proporcional ao seu grau, ou seja, quantas knowledge areas ele é similar (considerando similaridade maior que 0.2) e a grossura das arestas é proporcional à similaridade entre duas knowledge areas.

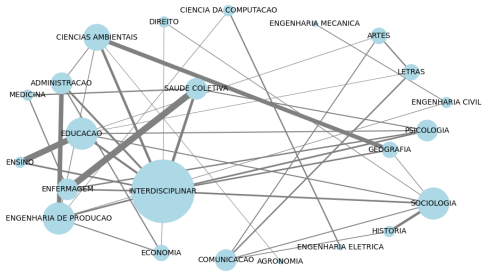

Figure 26: Grafo de similaridades

4.2 Itemsets frequentes

1. **Nova Matriz e Grafo de Similaridades**

   Com base nas keywords dos trabalhos, foram gerados conjuntos de itemsets frequentes para cada knowledge area, considerando um suporte mínimo de 0.001 (itemsets que ocorrem em pelo menos 0,1% dos trabalhos da área). A partir desses dataframes, uma nova matriz de similaridade foi criada, também com base no índice de Jaccard, avaliando a quantidade de itemsets iguais entre os trabalhos das knowledge areas.

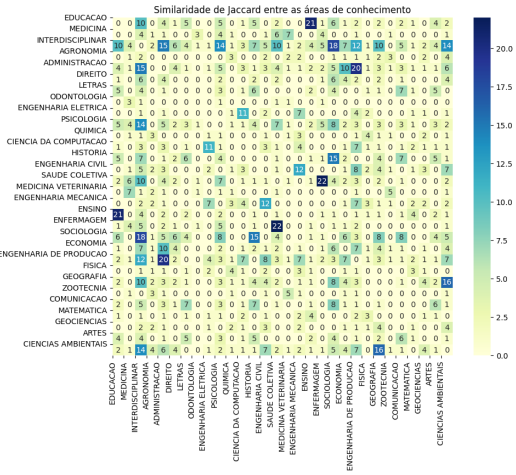

Figure 27: Matriz de similaridades

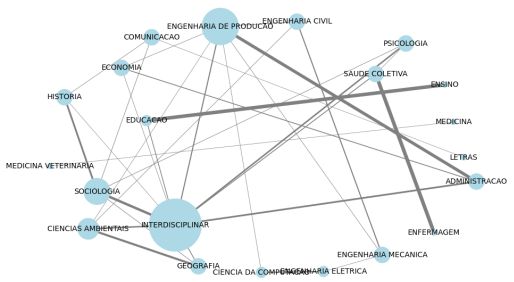

Figure 28: Grafo de similaridades

Mineração de Regras de Associação para cada área

Para cada knowledge area, foi feita uma mineração de itemsets frequentes nas keywords dos trabalhos correspondentes. Foi definido um suporte mínimo de 0.001. Essa etapa permitiu identificar combinações de keywords que aparecem com maior frequência nos trabalhos de cada área.

A partir da obtenção dos itemsets frequentes, foi realizada uma mineração de regras de associação entre as keywords de cada knowledge area. Nessa etapa, foi escolhido um suporte mínimo de 0.001 para gerar as regras de associação.

A métrica utilizada para avaliar as regras de associação foi a convicção, por ser uma ótima métrica para ordenar regras de associação, já que leva em conta o suporte dos itemsets.

Exemplos:

Regras de associação EDUCACAO:

	antecedents	consequents	conviction	confidence
5268	{INFANTIL, ESPEC}	{EDUCACA}	inf	1.00
8632	{POLI, MEDI, EDUCAC}	{ENSIN}	140.56	0.99
7412	{ADULT, DOCENT, EDUCACA}	{JOV}	94.62	0.99
7904	{POLI, CONTINU, PROFE}	{FORMACA}	93.48	0.99
7751	{DOCENT, PROFE, CONTINU}	{FORMACA}	90.17	0.99

Figure 29: Regras de associação em educação

Regras de associação CIENCIA DA COMPUTACAO:

	antecedents	consequents	conviction	confidence
2722	{APREND, NATUR, MAQUIN, PROCES}	{LINGU}	inf	1.0
2608	{NATUR, APREND, PROCES}	{LINGU}	inf	1.0
1958	{BANC, ORIENT}	{DAD}	inf	1.0
2691	{RED, NATUR, PROCES}	{LINGU}	inf	1.0
2678	{NATUR, MAQUIN, PROCES}	{LINGU}	inf	1.0

Figure 30: Regras de associação em ciência da computação

Regras de associação DIREITO:

	antecedents	consequents	conviction	confidence
6615	{FUNDAMENT, TRABALH, HUMAN}	{DIREIT}	inf	1.00
6223	{DEMOCRA, EST, FUNDAMENT}	{DIREIT}	inf	1.00
6601	{FUNDAMENT, SOC, HUMAN}	{DIREIT}	inf	1.00
2501	{DESENVOLV, MEI}	{AMBI}	113.71	0.99
6071	{NOV, CODIG, PROCES}	{CIVIL}	91.68	0.99

Figure 31: Regras de associação em direito

Mineração de Regras de Associação em todas as transações

Após a análise de similaridades, a mineração de regras de associação foi realizada para todo o conjunto de trabalhos de pós-graduação. Para isso, foi criada uma lista de keywords, composta pela união das 20 principais keywords de cada área. Em seguida, foram considerados apenas os trabalhos que continham keywords presentes nessa lista.

Devido ao desequilíbrio na quantidade de trabalhos por área, foi necessário estratificar o dataframe, gerando amostras homoegêneas do mesmo tamanho da quantidade de trabalhos da área com menos trabalhos.

Após a estratificação, as knowledge areas foram adicionadas como dummies no conjunto de dados, e foi possível obter regras de associação nas quais a knowledge area é antecedente ou consequente.

Vale ressaltar que, apesar de a ocorrência de alguns itemsets implicar na ocorrência de uma knowledge area, a regra de associação correspondente pode não ter confiança igual a 1 caso o itemset antecedente também seja antecedente de outras keywords.

	antecedents	consequents	conviction	confidence
3005	(SAUD, PRIM, KA_SAUDE COLETIVA)	{ATENCA}	135.491503	0.992683
590	(DENTIN)	{KA_ODONTOLOGIA}	54.737500	0.982340
2222	(SAUD, PRIM)	{ATENCA}	48.766223	0.979670
2990	(SAUD, PRIM, KA_ENFERMAGEM)	{ATENCA}	44.471427	0.977707
2751	(FAMIL, ESTRATEG)	{SAUD}	43.343001	0.978041

Figure 32: Regras de associação gerais

	antecedents	consequents	conviction	confidence
21100	{CORP, ALGEBR}	{KA_MATEMATICA}	inf	1.000000
21148	{ALGEBR, IDENT}	{KA_MATEMATICA}	inf	1.000000
21171	{NUMER, ALGEBR}	{KA_MATEMATICA}	25.455556	0.962025
49784	{FINIT, GRUP}	{KA_MATEMATICA}	23.200000	0.958333
57218	{NUMER, TEOREM}	{KA_MATEMATICA}	20.783333	0.953488

Figure 33: Regras de associação matemática

Análise de resultados Esta seção de mineração de regras de associação forneceu insights sobre padrões de itemsets em trabalhos de pós graduação. Foi possível observar a relação e similaridade entre knowledge areas através da matriz e do grafo de similaridades. Através das regras de associação, foram observadas as relações mais fortes

tanto entre as keywords e as knowledge areas quanto entre os próprias keywords.

5 Conclusão

Neste trabalho, nosso objetivo foi analisar um dataset sobre a pós-graduação no Brasil, com foco nas knowledge areas por meio das keywords. Identificamos que o banco de dados original apresentava diversos problemas, como datas inconsistentes, keywords com erros de digitação, dentre outrous. Além disso, notamos um grande desequilíbrio na distribuição de trabalhos por área de conhecimento.

Inicialmente, desenvolvemos um classificador utilizando uma rede neural para lidar com essas questões. Embora o desempenho não tenha sido significativamente melhor em comparação com modelos clássicos, o classificador aprendeu a criar uma representação dos trabalhos.

Em seguida, aplicamos o algoritmo K-means para realizar o agrupamento das knowledge areas, utilizando vetores médios como representação. O resultado foi a formação de 11 clusters que possuíam significado semântico, indicando que o modelo foi capaz de construir uma representação coerente das áreas.

Esses resultados reforçam a eficácia da abordagem adotada, na qual o modelo aprendeu a representar as knowledge areas com base nas keywords. Além disso, realizamos uma mineração dos dados em busca de regras de classificação, mesmo sem criar uma representação explícita das keywords. Essas regras permitiram analisar o comportamento geral das palavras-chave.

Como sugestões para trabalhos futuros, é importante realizar uma limpeza mais aprofundada do dataset, corrigindo os problemas identificados. Também seria interessante desenvolver uma rede neural mais robusta, explorando outros atributos do dataset para aprimorar a análise do agrupamento.

Em suma, este trabalho proporcionou insights valiosos sobre a pós-graduação no Brasil, fornecendo uma abordagem para analisar as knowledge areas com base em keywords. Esses insights podem ser futuramente aplicados para melhor agrupar as pós-graduações, visando mais a semática do trabalho do que categorias.