

Trabalho Prático Intermediário

INF 791 / INF 493: Tópicos Especiais II - Redes Complexas

Aluno: Lorena Souza Moreira

Matrícula: 108204

Professor: Júlio Cesar Soares dos Reis

13 de junho de 2025

Sumário

1	Coleta de Dados	2
1.1	Descrição da Rede e da Ferramenta de Coleta	2
1.2	O Grafo de Estudo e Suas Limitações	2
2	Análise da Rede Complexa	4
2.1	Breve Explicação da Rede Analisada	4
2.2	Distribuição de Grau	4
2.3	Componentes do Grafo	5
2.4	Coeficiente de Clusterização	5
2.5	Overlap da Vizinhança	6
2.6	Distâncias na Rede	7
2.7	Visualização do Grafo	7
3	Código Fonte	9

1 Coleta de Dados

1.1 Descrição da Rede e da Ferramenta de Coleta

Neste trabalho, foi utilizada uma rede baseada em um conjunto de dados públicos voltado para o estudo do discurso de ódio em português. A base de dados escolhida é proveniente do artigo "*A Hierarchically-Labeled Portuguese Hate Speech Dataset*", de Paula Fortuna e colaboradores (FORTUNA et al., 2019). Essa escolha se justifica por dois motivos principais. Em primeiro lugar, o português é uma das línguas mais faladas no mundo, mas ainda apresenta baixa representação em estudos quantitativos sobre discurso de ódio nas redes sociais. Em segundo lugar, o conjunto de dados em questão oferece uma estrutura rica, com anotações manuais e categorização binária, o que facilita a exploração e a modelagem de redes a partir de seu conteúdo textual.

O dataset foi construído a partir de mensagens coletadas no Twitter durante os dias 8 e 9 de março de 2017, totalizando 5.668 tweets escritos em português. O processo de coleta envolveu o uso da API de busca do Twitter, explorando duas estratégias principais. A primeira foi a utilização de palavras-chave e hashtags relacionadas ao discurso de ódio, como, por exemplo, “sapatão” ou a hashtag ‘#LugarDeMulherÉNaCozinha’. A segunda estratégia consistiu em mapear e extrair publicações de usuários previamente identificados como fontes recorrentes de conteúdo odioso. Essas abordagens complementares permitiram reunir um conjunto de dados mais representativo e diverso, abrangendo diferentes formas de discriminação, incluindo, mas não se limitando a, questões de religião, gênero, orientação sexual, etnia e migração.

Após a coleta inicial, os autores aplicaram filtros para eliminar mensagens duplicadas e conduziram uma amostragem que garantisse a heterogeneidade das fontes e dos tipos de discurso presentes no corpus. Como resultado, obteve-se um conjunto final com mensagens provenientes de 1.156 usuários distintos, representando uma variedade considerável de perfis e opiniões. O dataset foi então anotado manualmente, com classificação binária (discurso de ódio ou não), o que o torna particularmente útil tanto para tarefas de aprendizado supervisionado quanto para análises estruturais e semânticas em grafos.

Para este projeto, a rede foi construída a partir dos dados brutos contidos no arquivo CSV disponibilizado pelos autores no GitHub: <https://github.com/paulafortuna/Portuguese-Hate-Speech-Dataset/blob/master/2019-05-28_portuguese_hate_speech_binary_classification.csv>

1.2 O Grafo de Estudo e Suas Limitações

A partir do conjunto de dados com mensagens de discurso de ódio descrito anteriormente, foi construído um grafo com o objetivo de representar as relações léxicas entre palavras usadas nesse tipo de conteúdo. A construção do grafo foi realizada com base nos textos processados da coluna pro-text, onde cada mensagem foi previamente normalizada e tokenizada. A modelagem do grafo tem como base a coocorrência de palavras consecutivas, capturando padrões de associação direta no nível da estrutura sintática dos tweets.

A estrutura do grafo foi definida da seguinte forma:

- **Nódos (ou vértices):** cada nó do grafo representa uma palavra única presente nas mensagens classificadas como discurso de ódio. O nó armazena um atributo chamado count, que contabiliza o número de vezes em que essa palavra aparece no

corpus. Isso permite identificar termos recorrentes e avaliar sua importância relativa dentro da rede.

- **Arestas:** cada aresta conecta dois nós que aparecem lado a lado (coocorrência sequencial) em pelo menos uma mensagem. Por exemplo, se a sequência “mulher cozinha” aparece em um tweet, será criada uma aresta entre os nós "mulher" e "cozinha". As arestas são ponderadas com um atributo chamado *weight*, que representa o número total de vezes que a mesma coocorrência foi observada ao longo do corpus. Dessa forma, arestas com peso mais alto indicam pares de palavras que aparecem frequentemente juntas, sugerindo ligações semânticas ou discursivas mais fortes.

O grafo gerado é não direcionado, já que a ordem de aparecimento das palavras em um par não foi considerada relevante para esta análise inicial. Essa decisão permite tratar o grafo como uma rede de coocorrência simétrica, o que facilita a aplicação de algoritmos de detecção de comunidades e centralidade.

Quanto à abrangência do grafo, é importante ressaltar que ele representa apenas uma amostra limitada da totalidade de interações linguísticas no Twitter. O conjunto de dados original foi coletado apenas em dois dias (8 e 9 de março de 2017), utilizando a API de busca do Twitter, com base em palavras-chave e perfis previamente identificados. Essa abordagem de coleta induz um viés na seleção de mensagens, favorecendo conteúdos mais explícitos e ignorando contextos mais sutis ou indiretos de discurso de ódio. Além disso, a restrição temporal reduz a cobertura da coleta.

Consequentemente, o grafo obtido é apenas uma **representação parcial** da rede semântica do discurso de ódio em português. Ele não inclui outras formas de interação entre usuários, como retweets, menções ou respostas, nem leva em consideração o contexto conversacional em que os tweets foram emitidos. Mesmo com essas limitações, o grafo fornece uma base relevante para análise exploratória, permitindo observar padrões lexicais dominantes, termos centrais no vocabulário ofensivo e potenciais agrupamentos temáticos com base na proximidade entre palavras.

2 Análise da Rede Complexa

2.1 Breve Explicação da Rede Analisada

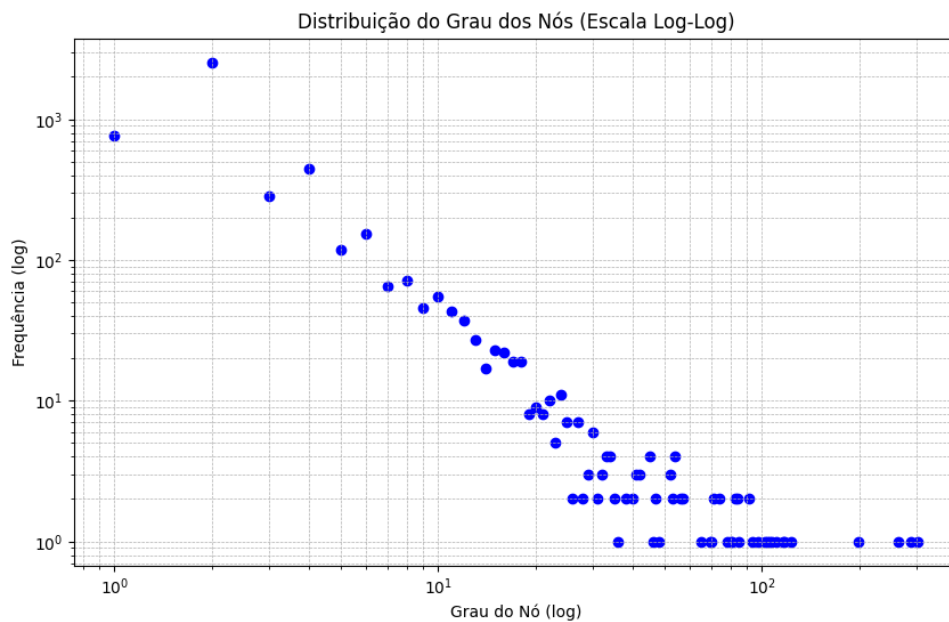
A rede analisada foi construída a partir de mensagens com discurso de ódio em português extraídas do Twitter. Trata-se de um grafo não direcionado e ponderado, no qual os **nós representam palavras únicas** presentes nos tweets, acompanhadas de um atributo que indica sua frequência no corpus. As **arestas conectam palavras que aparecem lado a lado** nas mensagens, sendo ponderadas pelo número de vezes em que essa coocorrência ocorre. Essa estrutura permite identificar padrões lexicais e relações

2.2 Distribuição de Grau

Observa-se uma frequência muito alta de nós com grau baixo, especificamente, 764 nós com grau 1 e 2541 nós com grau 2 e poucos nós com grau muito alto (por exemplo, graus 198, 264, 287 e 302). Quando plotada em escala log-log, a distribuição de grau forma uma linha aproximadamente reta, característica de uma lei de potência.

O grau médio calculado para a rede é 4,52, o que é relativamente baixo considerando que alguns nós possuem mais de 300 conexões. Isso indica que a rede é **esparsa**, ou seja, o número de conexões presentes é muito menor do que o número total possível, o que é esperado em redes de coocorrência de palavras, onde as conexões não são aleatórias, mas estruturadas de forma específica. A rede apresenta hubs evidentes, nós com grau excepcionalmente alto em comparação com a média. Esses hubs funcionam como pontos centrais que conectam grupos de palavras que, de outra forma, estariam desconectados.

Em resumo, a rede analisada tem estrutura baseada em uma lei de potência, um baixo grau médio e a presença marcante de hubs. No contexto do grafo de palavras, isso significa que a maioria das palavras está conectada a poucas outras, enquanto poucas palavras atuam como "coladores" do vocabulário, sendo essenciais para a coesão da rede.



2.3 Componentes do Grafo

O grafo analisado possui 10 componentes conexos distintos, conforme apresentado na Tabela 1. Destaca-se a presença de um componente gigante, o Componente 1, que contém 4873 nós e 11056 arestas, concentrando a grande maioria dos elementos da rede. Os demais componentes são significativamente menores, com tamanhos variando entre 2 e 7 nós.

A existência de um componente gigante indica que a maior parte dos nós está fortemente conectada, formando uma estrutura coesa que permite o fluxo de informações ou relações entre seus elementos. Já os múltiplos componentes menores representam subgrafos isolados ou pouco conectados, que podem corresponder a grupos específicos ou palavras que não se relacionam diretamente com o núcleo principal da rede, ou seja, pouco usuais.

Componente ID	Nº de Nós	Nº de Arestas
1	4873	11056
2	7	6
3	7	6
4	7	6
5	4	3
6	3	2
7	2	1
8	2	1
9	2	1
10	2	1

Tabela 1: Distribuição do tamanho dos componentes do grafo.

2.4 Coeficiente de Clusterização

O coeficiente de clusterização global (médio) do grafo é 0,0512. Em uma escala absoluta de 0 a 1, esse valor pode parecer baixo, mas sua real significância aparece quando o comparamos com o valor esperado para uma rede aleatória com o mesmo número de nós e grau médio.

No caso de uma rede aleatória (modelo de Erdős-Rényi), o coeficiente de clusterização esperado (C_{rand}) pode ser estimado como:

$$C_{\text{rand}} \approx \frac{\langle k \rangle}{N}$$

Com grau médio $\langle k \rangle = 4,52$ e aproximadamente $N = 4900$ nós, o valor esperado seria cerca de 0,001. Isso significa que o coeficiente observado é aproximadamente 50 vezes maior do que o esperado para uma rede aleatória. Tal diferença indica uma estrutura não aleatória, com forte tendência à formação de *clusters* ou "panelinhas".

Além disso, a distribuição dos coeficientes de clusterização locais revela a presença de muitos nós com valor igual a zero, indicando que esses nós são periféricos e possuem poucas conexões entre seus vizinhos. Esses elementos frequentemente atuam como pontes locais, conectando diferentes palavras ou grupos que, de outra forma, permaneceriam isolados.

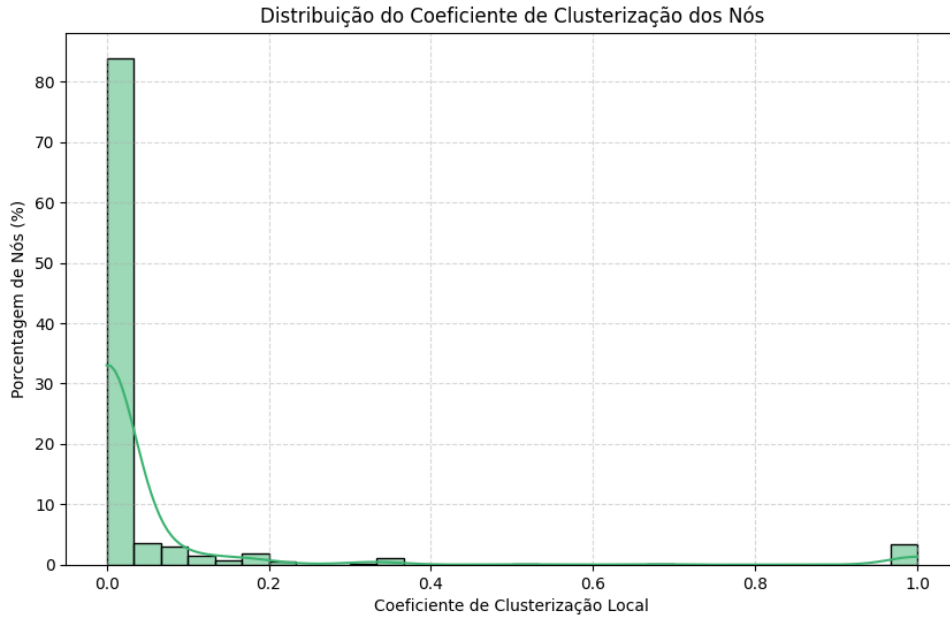


Figura 2: Distribuição do coeficiente de clusterização dos nós.
Coeficiente de clusterização global (C): 0.0512.

2.5 Overlap da Vizinhança

A predominância de valores iguais a 0,0 no Coeficiente de Jaccard para a maioria dos pares de nós não conectados indica que esses pares compartilham poucos ou nenhum vizinho em comum. Essa característica sugere que a rede apresenta uma estrutura modular ou fragmentada, na qual os nós estão organizados em grupos relativamente isolados, com baixa sobreposição entre as vizinhanças. Em contextos como grafos de coocorrência de palavras, isso revela que a maioria das palavras ocorre em contextos distintos, o que reforça a especificidade e a dispersão da rede. Além disso, essa baixa similaridade local entre nós implica que a probabilidade de formação de novas conexões entre esses pares é pequena, evidenciando uma rede estruturalmente esparsa e altamente segmentada.

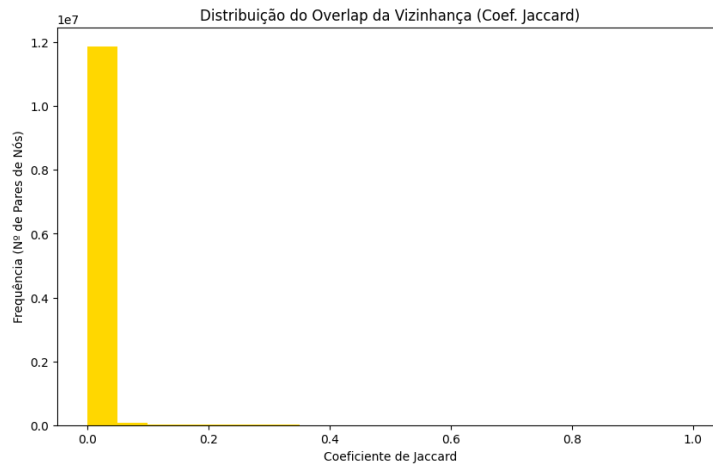


Figura 3: Distribuição do overlap da vizinhança.

2.6 Distâncias na Rede

A distância média entre os nós no maior componente conectado da rede é 4,35, indicando que, em média, é necessário percorrer pouco mais de quatro arestas para ir de um nó a outro dentro desse componente. Essa medida sugere que a rede possui uma boa conectividade global, permitindo a comunicação eficiente entre os elementos. A distribuição das distâncias apresenta um padrão típico de redes complexas: a maior parte dos pares de nós está separada por pequenas distâncias, enquanto poucas conexões ocorrem em distâncias maiores, decrescendo gradativamente até a distância máxima observada de 18. Essa distribuição confirma que a rede exibe o fenômeno conhecido como "mundo pequeno" (small-world), caracterizado por caminhos curtos entre a maioria dos nós, mesmo em redes de grande escala.

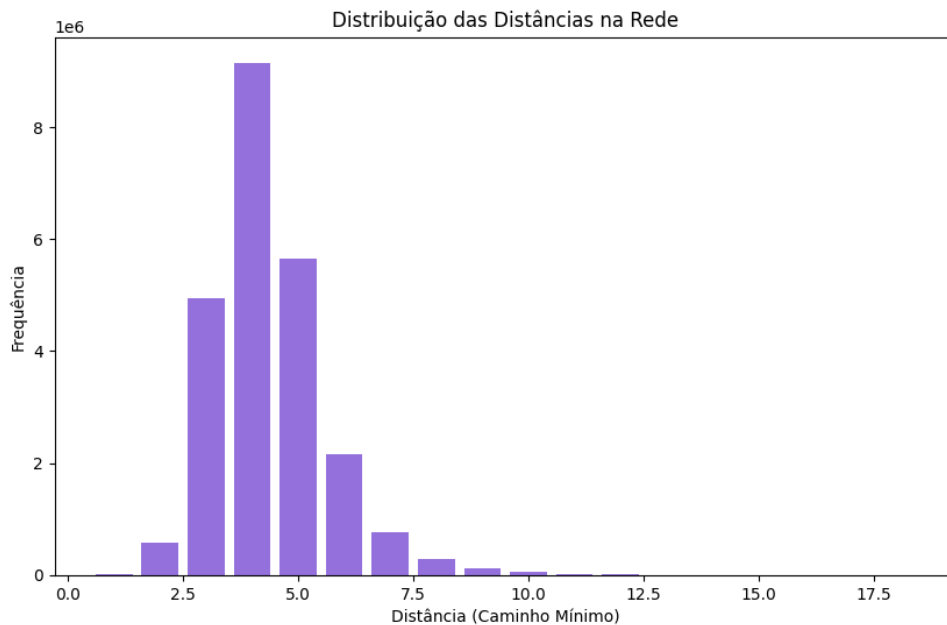


Figura 4: Distribuição das distâncias entre os nós da rede.
Distância média no maior componente conectado (l): 4.35.

2.7 Visualização do Grafo

O grafo correspondente ao maior componente conexo é grande demais para permitir uma análise visual eficaz, enquanto os demais componentes são extremamente pequenos e irrelevantes. Por isso, a visualização direta do grafo não é a abordagem mais adequada para compreender sua estrutura.

8

3 Código Fonte

Todo o código-fonte está disponível em um repositório público no GitHub.

- **Repositório (preferencial):** github.com/LorenaSouzaMoreiraa/RedesComplexas-TpIntermediario
- **Anexo (alternativo):** Também incluímos um arquivo compactado com o código no envio do trabalho.

Referências

FORTUNA, P.; SILVA, J. R. da; WANNER, L.; NUNES, S. et al. A hierarchically-labeled portuguese hate speech dataset. In: *Proceedings of the third workshop on abusive language online*. [S.l.: s.n.], 2019. p. 94–104.