Avaliação de redação com base em propriedades topológicas de redes complexas

Felipe F. R. Melo, Thales M. Leijoto

Ciência da Computação – Universidade Federal do São João Del-Rei (UFSJ)

felipefrmelo@hotmail.com, thalesmradl@gmail.com

Abstract. This work aims to describe and discuss the results of essay analysis experiments when they are modeled as complex networks. You experiments did not indicate a correlation as expressive as expected between the topographic parameters of the complex networks and the evaluations of the essays given by specialists. However, it is believed that a good calibration of parameters and with more appropriate natural language processing methods, this proposal can be useful to assess the quality of essays. Modeling in complex networks is independent of language and simplifies the work of automatic text analysis, showing a promising alternative to linguistically motivated methods.

Palavras-chave: Complex networks, argumentative-essay texts.

Resumo. Este trabalho tem como objetivo descrever e discutir os resultados de experimentos de análise de redações quando estes são modelados como redes complexas. Os experimentos não indicaram uma correlação tão expressiva quanto se esperava entre os parâmetros topográficos das redes complexas e as avaliações das redações dadas por especialistas. Contudo, acredita-se que com uma boa calibragem de parâmetros e com métodos de processamento de linguagem natural mais adequados, esta proposta pode ser tornar útil para avaliar a qualidade de redações. A modelagem em redes complexas é independente de língua e simplifica o trabalho de análise automática de textos, mostrando-se uma alternativa promissora aos métodos linguísticamente motivados.

Palavras-chave: Redes complexas, textos argumentativo-dissertativos.

1. Introdução

Os sistemas de Processamento de Língua Natural (PLN), até a década de 1990, seguiam o formalismo conhecido como simbólico, no qual todo o conhecimento deve ser expresso em gramáticas e em léxicos, fazendo com que todas as possíveis situações devam ser previstas durante a fase de desenvolvimento para que possam ser resolvidas pelo sistema. Entretanto, o que se verifica na prática é uma queda no índice de sucesso desses sistemas na medida em que eles se tornam mais abrangentes e robustos. Recentemente, a área de PLN passou a estudar e a aplicar novos formalismos, devido principalmente às limitações do paradigma simbólico, mas também devido a um grande obstáculo que vem impulsionando os pesquisadores da área na busca por outras abordagens até os dias de hoje: a análise profunda de um texto visando à compreensão de seu significado. Entre os modelos que passaram a ser utilizados, destacam-se os estatísticos e os conexionistas, provenientes dos estudos em Aprendizagem Automática. Posteriormente, o uso dessas novas técnicas proporcionou a construção, com sucesso, de alguns sistemas de PLN a partir de *corpora* de textos reais, dispensando o conhecimento linguístico explícito indispensável na abordagem simbólica. Os melhores etiquetadores (Part-of-Speech Taggers) disponíveis atualmente são resultados do uso de métodos estatísticos de aprendizado, fato que reforça a tendência em buscar novas alternativas na análise automática de línguas naturais. Nesse contexto, uma opção a ser considerada é a aplicação de redes complexas no PLN, um conceito recente, proveniente da mecânica estatística e que faz uso intenso da teoria dos grafos, ainda pouco estudado no âmbito do processamento automático de línguas naturais.

Anteriormente, os grafos randômicos eram os mais utilizados para estudar sistemas modelados como redes. Mas o interesse em sistemas complexos por parte dos cientistas estimulou uma reconsideração desse paradigma de modelagem. Como é cada vez mais evidente que a topologia e a evolução dessas redes são governadas por princípios robustos de organização (e não simplesmente randômicos, daí o nome "redes complexas"), sentiu-se a necessidade de desenvolver ferramentas para capturar quantitativamente esses princípios. Grande parte das recentes descobertas está relacionada à maneira como as redes do mundo real diferem das redes randômicas. Existe também uma crescente necessidade de entender o comportamento do sistema como um todo, movendo-se para além das abordagens reducionistas.

A linguagem humana também pode ser entendida como uma rede complexa. Cancho e Solé (2001) apresentam a análise de uma rede derivada do British National Corpus, sendo que os nós dessa rede representam as palavras, e suas arestas conectam palavras que aparecem no córpus pelo menos uma vez, em sequência ou separadas por uma palavra. Essa rede contém 478.773 nós e 1,77 x 107 arestas. Outra rede foi construída, semelhante à anterior, com a diferença de que apenas são considerados os pares de palavras consecutivas (i,j) que ocorrem mais vezes do que seria esperado quando a independência entre as palavras é assumida, ou seja, quando $p_{ij} > p_i p_j$. Essa

rede apresenta 460.902 nós e 1,61 x 107 arestas. Foi mostrado que as duas redes apresentam as características small-world e livre de escala, indicando que essa rede de palavras pertence à mesma classe, por exemplo, da Internet e da World Wide Web. Outro exemplo de estudo em linguagem natural envolvendo redes complexas foi desenvolvido por Sigman e Cecchi (2002). Nesse trabalho, o banco de dados Wordnet [Miller 1985] foi mapeado em uma rede, no qual os nós representam os diferentes significados das palavras (neste caso, apenas dos substantivos) e as arestas refletem as relações semânticas (como antonímia e hipernímia). É mostrado que a polissemia é responsável pela existência de hubs, os quais deixam os conceitos mais próximos entre si, tornando a rede small-world. Em outro estudo [Motter et al. 2002], um thesaurus da língua inglesa é modelado de modo que, na rede, duas palavras estão conectadas se representam conceitos similares. Essa rede apresentou a propriedade small-world e livre de escala (assintoticamente). Já Costa (2003) aplicou um modelo de rede em um experimento psicofísico, no qual uma pessoa fornece livremente uma palavra que julgue relacionada a outra palavra apresentada por um computador. Esse processo é repetido diversas vezes, e as arestas são criadas entre as palavras associadas pela pessoa. A distribuição dos graus de saída dessa rede indica que ela é uma rede livre de escala.

Nesse contexto de estudos recentes que evidenciam propriedades de redes complexas nas redes inspiradas na linguagem humana, este trabalho investiga sua potencial utilidade na análise de textos. Por meio do uso de medidas estatísticas independentes de língua, nosso objetivo é avaliar automaticamente critérios como qualidade, legibilidade, coerência, etc. Este trabalho busca direcionar para uma possível aplicação resultante do encontro entre o processamento automático de línguas, uma área multidisciplinar por excelência, e a efervescente área chamada redes complexas.

2. Metodologia

2.1. Extração dos dados

Os dados utilizados neste trabalho foram retirados do banco de redações do portal de educação Brasil Escola. Esse banco possui mais de 6 mil redações, de diversos temas, corrigidas por especialistas de acordo com os critérios do ENEM. As redações são escritas e enviadas por internautas anônimos mensalmente com tema definido pelo Brasil Escola.

Esses dados foram extraídos por um coletor de dados desenvolvido por Guilherme Passero, que disponibilizou, em seu github (*gpassero*), o código fonte e um arquivo json já contendo mais de 6 mil entradas extraídas pelo seu coletor até o ano de 2017.

2.1.1. O que são os dados

Vários dados foram coletados do site, como: título, data, comentários das redações, url, descrição, etc.

Porém os dados utilizados para a realização deste trabalho foram: o texto da redação em si, a nota total dada pelos especialistas e as notas parciais para cada critério exigido pelo ENEM.

Os critérios de avaliação dividem-se em 5:

1. Demonstrar domínio da modalidade escrita formal da língua portuguesa;
2. Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa;
3. Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista;
4. Demonstrar conhecimento dos mecanismos linguísticos necessários para a argumentação;
5. Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Cada um dos 5 critérios recebeu uma nota de 0 a 200 pelos especialistas, resultando em notas finais entre 0 a 1000.

2.2. Pré-processamento do texto

A modelagem utilizada neste trabalho requer um tratamento dos textos antes de representá-los como redes complexas. O objetivo dessa representação por redes complexas é codificar as relações entre os conceitos dos textos. Para tanto, cada texto tem suas stopwords removidas, eliminando assim as palavras com pouco significado. Além disso, as palavras restantes são lematizadas, a fim de agrupar conceitos de mesma forma canônica, mas com flexões diferentes.

2.3. Modelagem da rede

Após o pré-processamento do texto, as N palavras distintas passam a representar os nós da rede, e a sequência de palavras resultante é utilizada na criação das arestas, de modo que, para cada par de palavras consecutivas, existe uma aresta direcionada correspondente na rede. As arestas apresentam pesos, os quais indicam o número de vezes que as respectivas associações de palavras aparecem no texto. Note que, devido à etapa anterior de lematização, as palavras, por exemplo, “aparecem” e “aparecerão” serão representadas por apenas um nó na rede, neste caso rotulado por “aparec”. A

Figura 1 mostra o poema “No meio do caminho”, de Carlos Drummond de Andrade, junto ao resultado caso fosse modelado como um grafo.

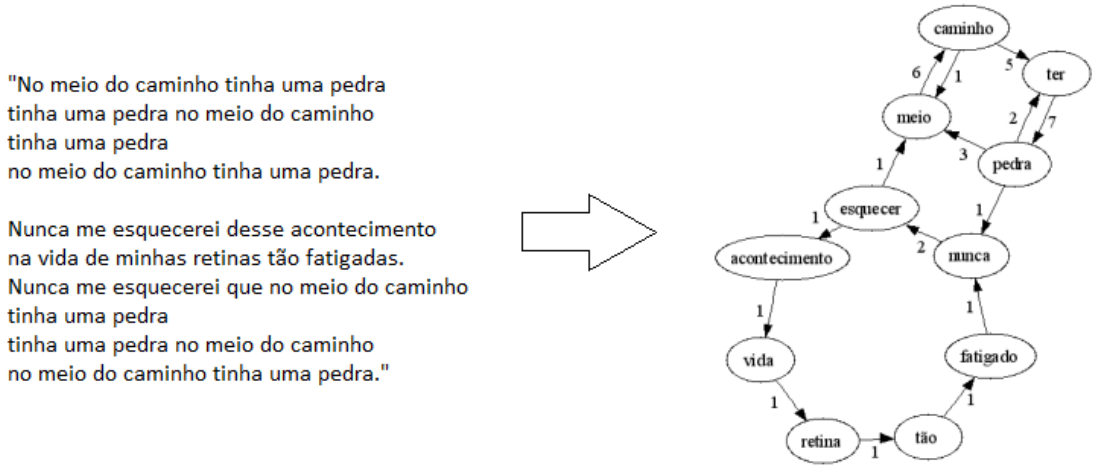

Figura 1. Poema “No meio do caminho”, de Carlos Drummond de Andrade, junto à sua modelagem.

2.4. Experimentos

Para cada rede foram computadas a quantidade de vértices e arestas, a média dos graus de saída (como em um dígrafo a média dos graus de entrada é igual à dos graus de saída, foi calculado somente o grau de saída), a média do coeficiente de aglomeração para todos os nós, a média dos caminhos mínimos entre todos os pares de nós da rede sem considerar o peso das arestas e também considerando o inverso do peso para cada aresta, com o intuito de dar ênfase às associações de palavras com maior peso. Foi calculado também o grau de assortatividade da rede, a densidade, e as métricas de centralidade de grau, de proximidade (closeness) e de intermediação (betweenness). Por fim, foi também executado o algoritmo de pagerank na rede.

A fim de identificar padrões que correlacionam as propriedades topológicas intrínsecas das redes, que por sua vez representam redações, e suas respectivas notas quanto a certos critérios, os resultados obtidos das métricas calculadas obtidos foram agrupados, estruturados e expostos de maneira gráfica para auxiliar na análise.

3. Discussões

A Figura 2 e a Figura 3 mostram exemplos de redes geradas por uma redação de nota 100, e uma redação de nota 1000, respectivamente.

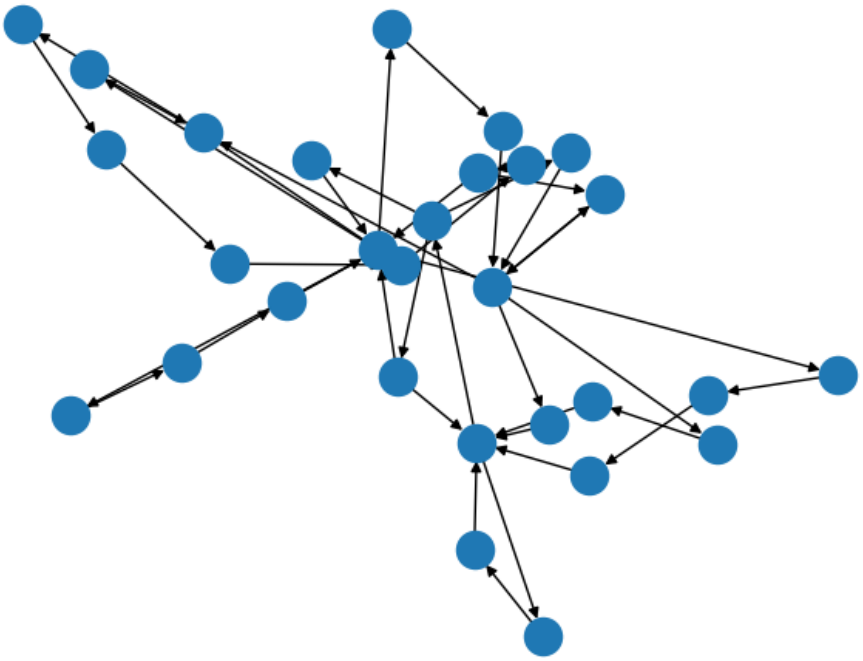

Figura 2. Grafo gerado por uma redação com nota final 100.

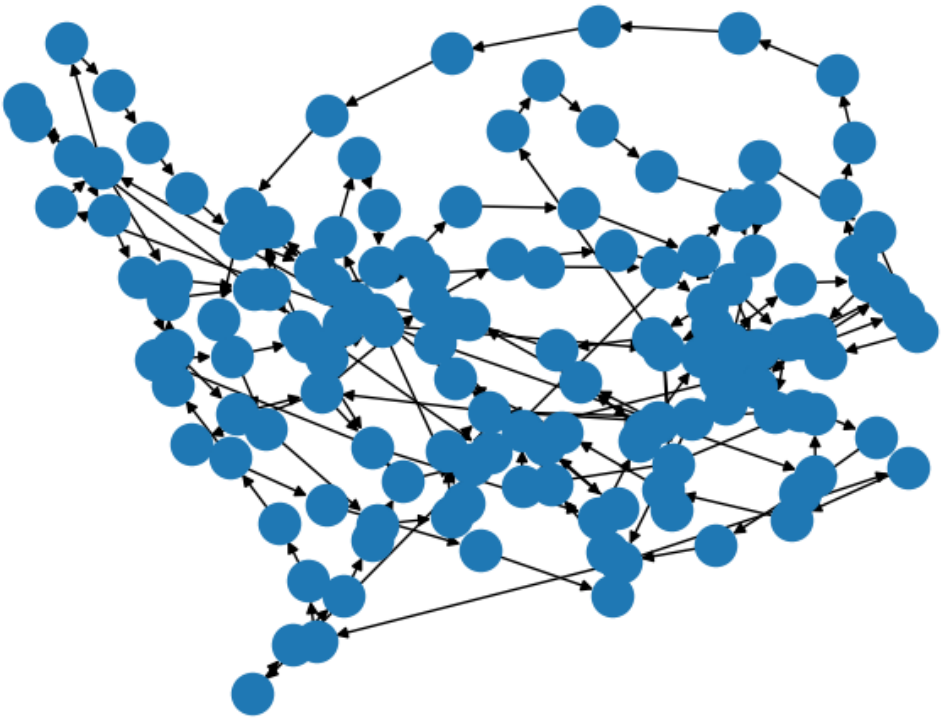

Figura 3. Grafo gerado por uma redação com nota final 1000.

Com base nestas duas figuras, superficialmente, já conseguimos de antemão observar algumas coisas, a respeito das características das redes em relação a nota final. Como por exemplo a quantidade de vértices e arestas, a redação com nota 1000 possui relativamente mais vértices que a redação com nota 100.

Antes de entrar de fato na análise e discussão dos resultados, um adendo deve ser feito: como explicado anteriormente (seção 2.1), são atribuídas cinco notas para cada redação, uma para cada competência e a soma das notas de competência é a nota final. Cada competência avalia certo aspecto da redação. Portanto, não se espera que as propriedades topológicas da redação modelada em grafos correlacione com todas as cinco competências. Espera-se que a competência 2 tenha uma relação mais forte com as propriedades da rede, pois se trata de uma competência voltada ao compreendimento da estrutura textual dissertativa-argumentativa e as inter-relações entre palavras e frases. Por este motivo, vamos focar na discussão desta competência. Além da à nota final, para verificar se pode-se tirar alguma conclusão, no geral, a respeito da redação a partir de suas características.

A Figura 4 mostra graficamente a relação entre a competência 2 e cada propriedade extraída das redes complexas.

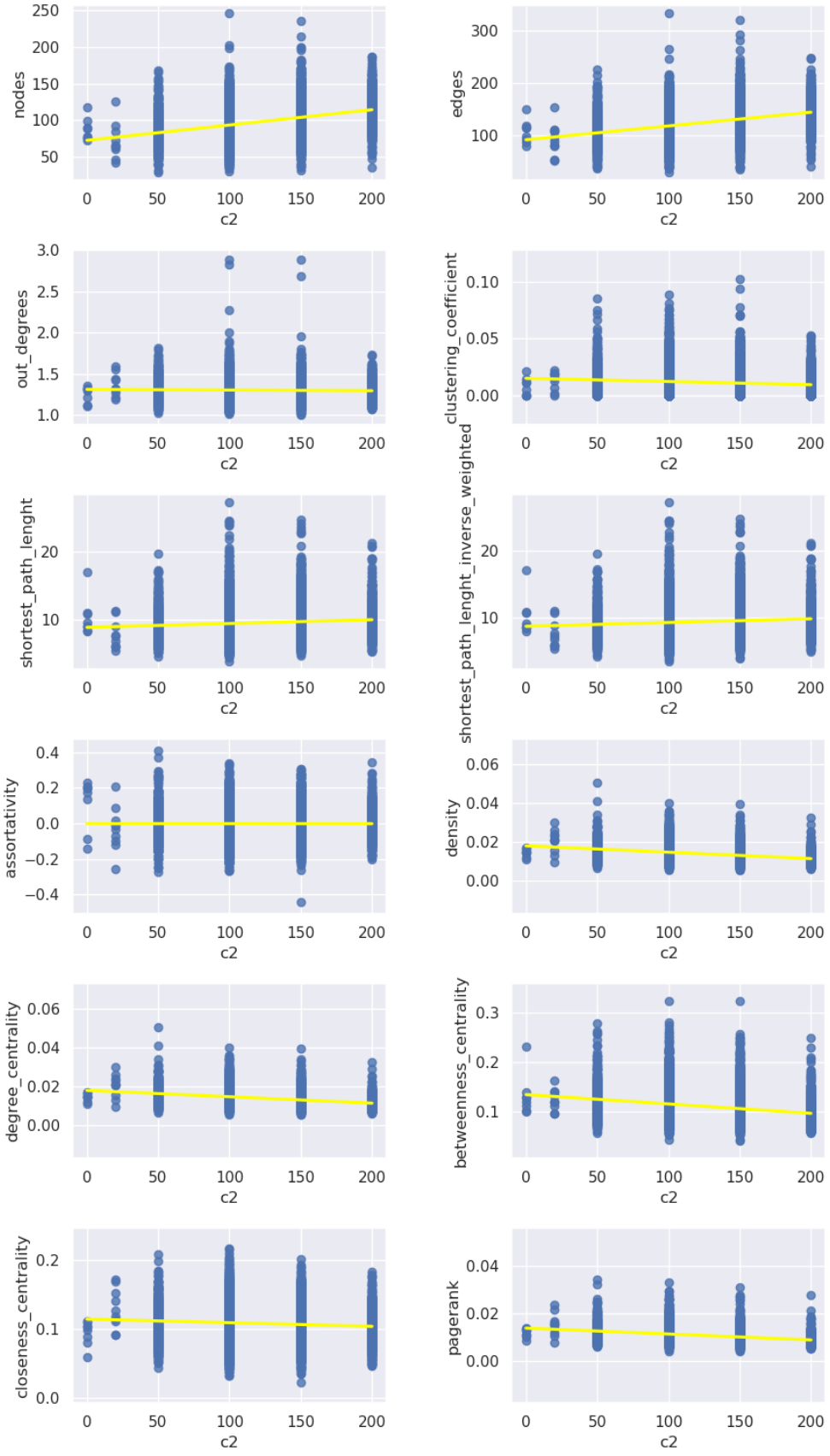

Figura 4. Relação entre a competência 2 e as propriedades topológicas extraídas das redes.

De acordo com a Figura 4, confirmamos o que foi observado nos grafos de nota 100 e nota 1000 apresentados anteriormente, a competência 2 tende a ser maior à medida que aumenta o número de nós e arestas. Isso pode indicar que redações com textos com maior variedade de palavras tendem a obter uma melhor avaliação no critério vigente da competência 2, que é a correto uso da estrutura dissertativa-argumentativa. Redações com um conjunto menor de palavras distintas, seja porque a redação tenha um número reduzido de palavras no geral ou seja porque é grande a ocorrência de palavras iguais obtiveram notas inferiores.

Ainda sobre a Figura 4, para a densidade, centralidade de grau, centralidade de intermediação, e pagerank é possível notar que à medida que estas métricas diminuem o de valor, a nota na competência 2 tende a aumentar. Traduzindo essas medidas para as redações, temos que um vasto vocabulário, juntamente com a utilização de palavras chaves contribui para essas medidas terem um menor valor, fazendo assim o texto ter uma nota maior.

Já para as medidas de caminho mínimo, baseado no Figura 4, nota-se uma pequena tendência de aumentar a nota à medida que aumenta a média da distância mínima entre nós. Isso mostra a habilidade de bons escritores de conseguir manipular conexões longas sem que o texto tenha sua qualidade prejudicada.

Por fim, a respeito da Figura 4, para o grau de saída, coeficiente de agrupamento, grau de assortatividade e centralidade de proximidade a correlação com a competência 2 a quase que mínima. Para a competência, medidas que tendem a mensurar uma correlação e um agrupamento entre os nós não é eficiente, pois essa competência o desenvolvimento da estrutura do texto e o domínio da estruturação requerida, o que não pode ser visto por essas medidas.

A Figura 5 mostra graficamente a relação entre as notas finais e cada propriedade extraída das redes complexas.

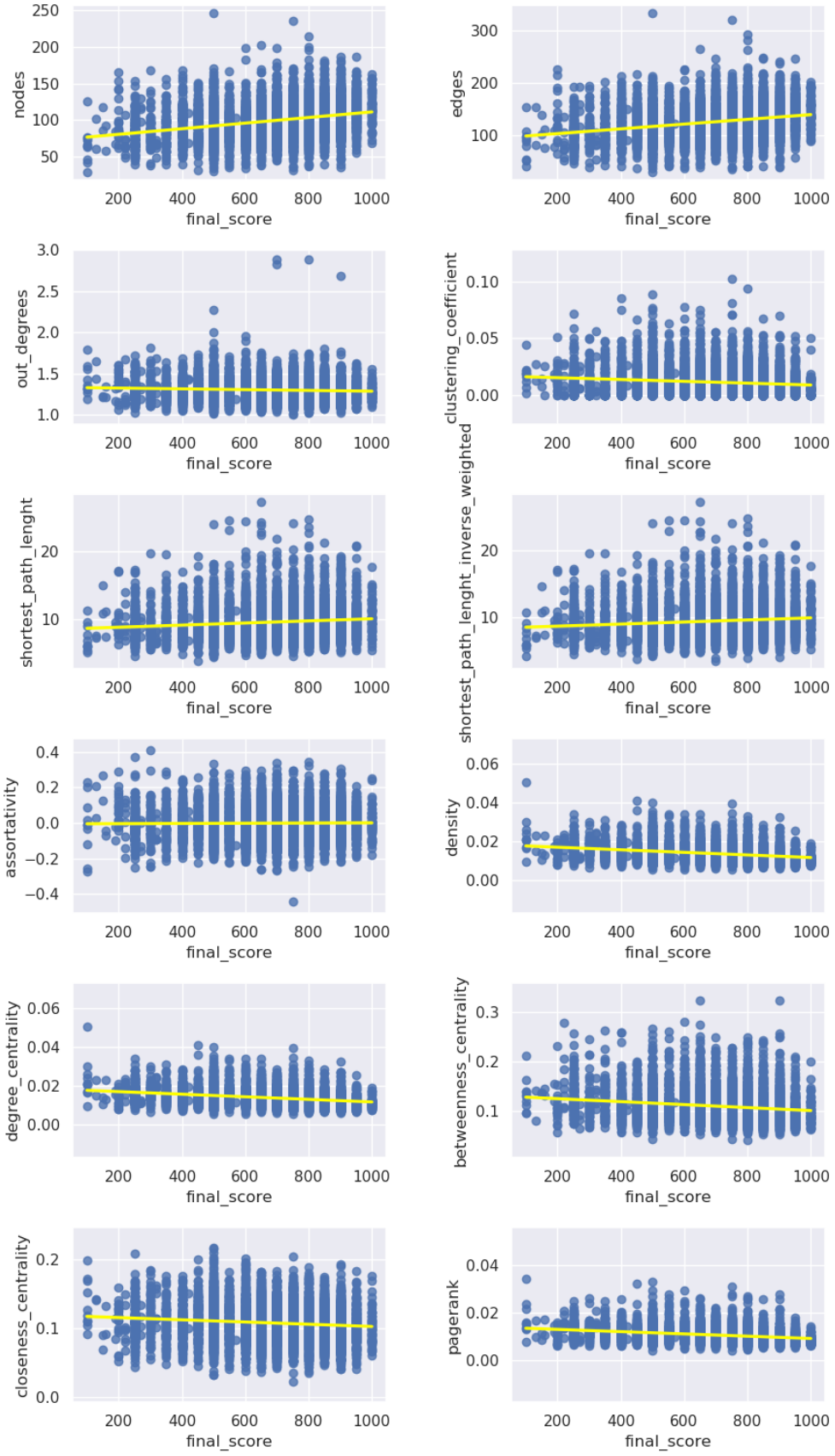

Figura 5. Relação entre a nota final e as propriedades topológicas extraídas das redes.

Observando a Figura 5, podemos perceber que a nota final também mantém uma relação diretamente proporcional ao número de nós e arestas. Isso demonstra que a utilização de uma diversidade maior de palavras está atrelada a redações com notas maiores.

Podemos ver também que as distâncias dos caminhos mais curtos e dos caminhos mais curtos com pesos crescem juntamente com as notas. Mostrando que redações com boas notas tendem conexões longas de palavras distintas.

Quanto às métricas de densidade, coeficiente de agrupamento, centralidade de grau, centralidade de intermediação, e pagerank, assim como na competência 2, podemos ver que tem valores menores conforme as notas aumentam. Isso pode ser explicado vendo que boas redações evitam repetições e tem um vocabulário mais extenso, formando assim uma rede mais esparsa e com poucos agrupamentos. Um coeficiente de agrupamento acentuado em notas baixas pode ser reflexo argumentos de caráter cíclico e repetitivo.

Para o restante das medidas, a partir da Figura 5, não foi possível observar nenhum comportamento tendencioso.

As figuras mostradas anteriormente, mostravam todas as redações e suas respectivas notas junto com os seus valores nas propriedades calculadas, além da linha de regressão linear que auxilia na projeção de valores (valor esperado) de uma variável y, dados os valores de x. A Figura 6 que será apresentada agora mostra esta relação por outra métrica, será considerado a média das redações de um dado intervalo de notas para todas as métricas calculadas.

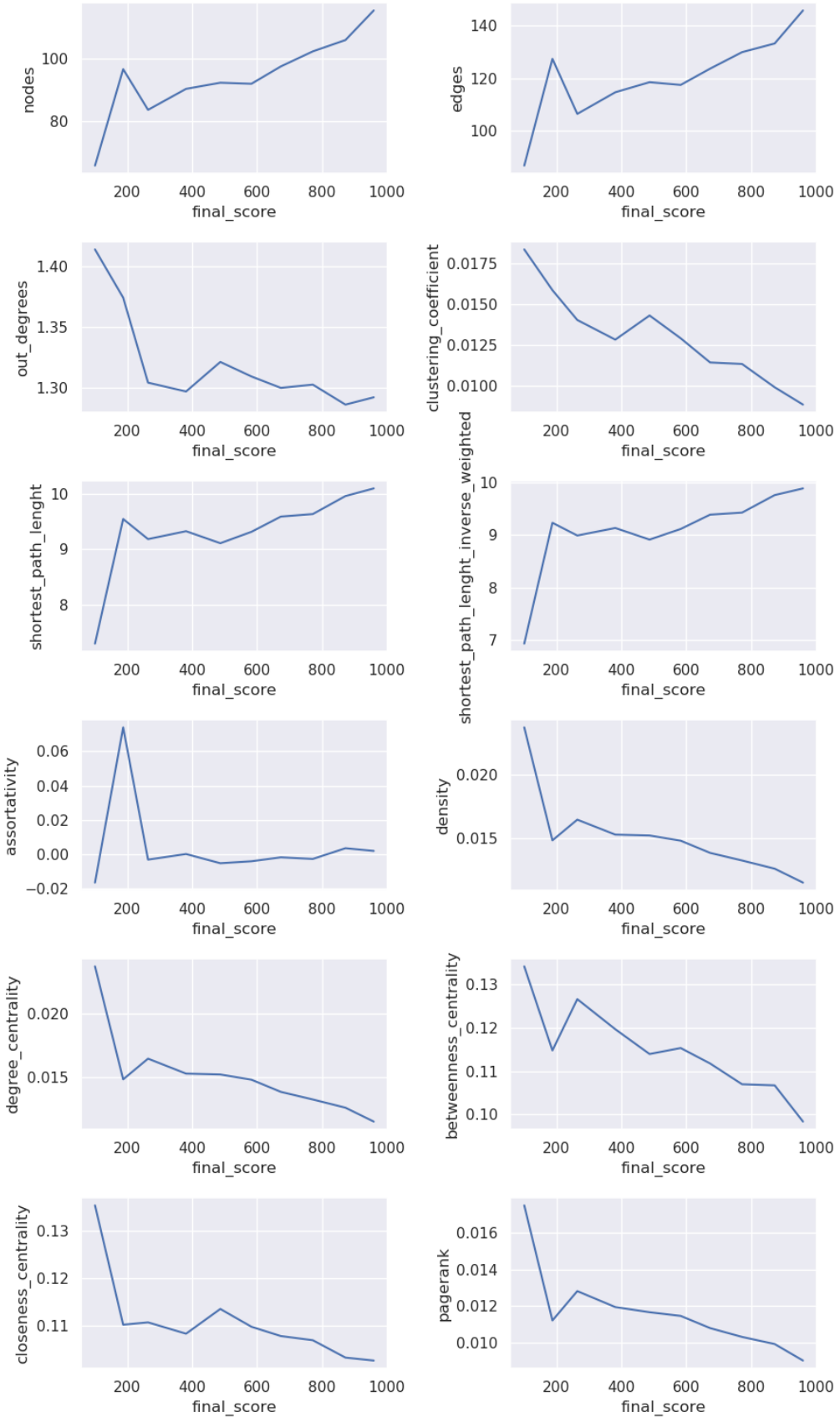

Figura 6. Valores médios entre a nota final e as propriedades topológicas extraídas das redes.

No intervalo de 0 a 200 no eixo x da Figura 6, para todas as métricas, observou-se uma mudança brusca do valor de f (x), isso ocorreu pois para este intervalo na base de dados usada não existem muitas redações com estas nota (aproximadamente 100 para um total de mais de 6000 redações), por isso não será levado em consideração as informações obtidas para as notas menores de 200.

Na Figura 6, conseguimos, novamente, observar o aumento da nota final à medida que aumenta o número de nós, arestas, a distância do caminho mínimo, por motivos que já foram explicados anteriormente. Também observamos um aumento na nota final à medida que diminui densidade, centralidade de grau, centralidade de intermediação, e pagerank. E por fim, notamos que as métricas de grau de saída, assortatividade e centralidade de aproximação variam mas sem causar notável queda ou crescimento nas notas finais.

A Figura 7 resume as correlações entre todas as propriedades de redes calculadas e todas as notas das redações.

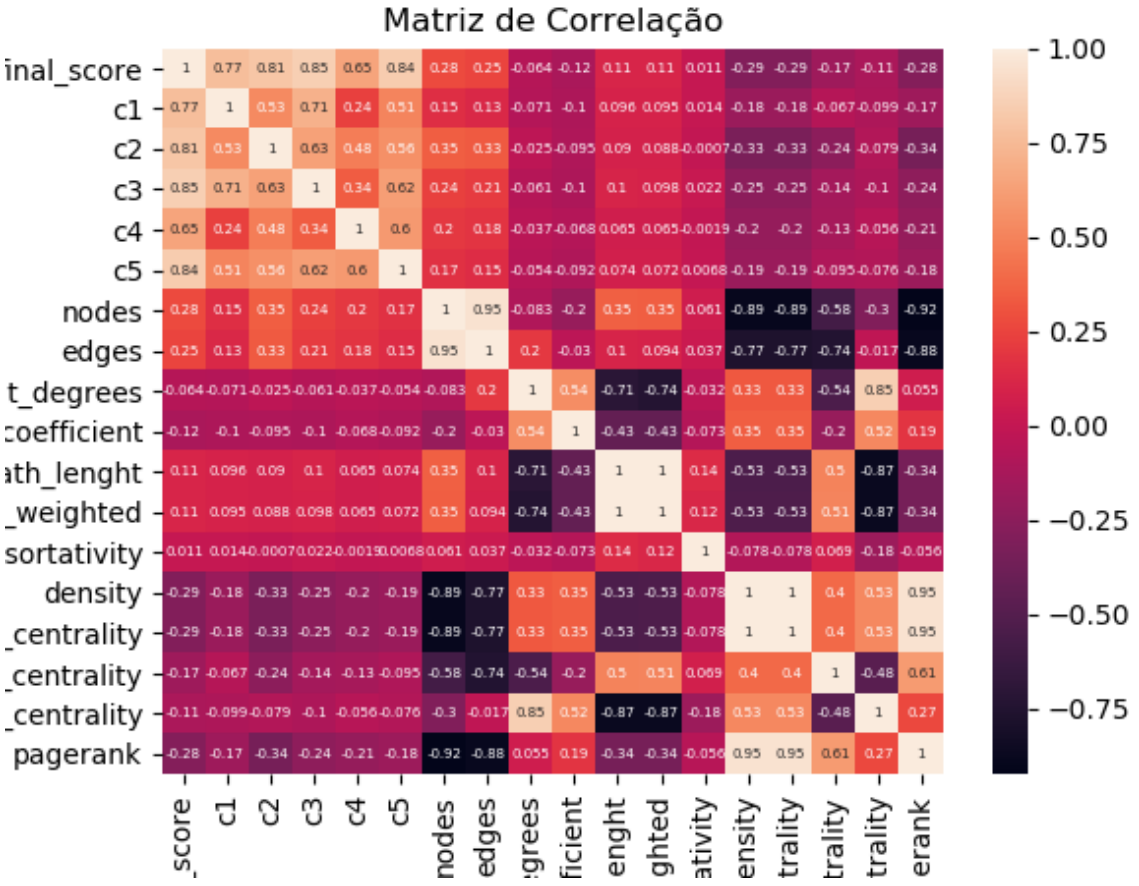

Figura 7. Matriz de coeficiente de correlação.

De acordo com a Figura 7, reforçamos algumas coisas que já foram ditas, como por exemplo a correlação positiva entre o número de nós e arestas com a competência 2 e com a nota final e a correlação negativa entre a densidade, centralidade de grau e

centralidade de intermediação com a competência 2 e com a nota final. Contudo, no geral, nota-se que as correlações entre as notas e as propriedades das redes não são tão expressivas.

4. Conclusões

Este trabalho ressalta a utilização dos novos conceitos de redes complexas na análise de redações com notas atribuídas por juízes humanos. As medidas estatísticas calculadas das redes foram confrontadas com as notas atribuídas às redações, indicando que, embora não tão expressivos os coeficientes de correlações, as propriedades de uma rede conseguem expressar algumas coisas a respeito de uma redação de caráter dissertativa-argumentativa.

Desejava-se apresentar um modelo preditivo de notas a partir das propriedades extraídas das redes, porém devido aos baixos fatores de correlações entre a maioria das propriedades da rede e as notas, que por sua vez pode ter sido ocasionada devido a quantidade de redações na base de dados ser insuficiente para treinar o modelo com todas as features desejadas, o desenvolvimento deste modelo não foi continuado. Acredita-se que com um pré-processamento mais intenso sobre os textos das redações, como a etiquetagem morfossintática, ou até mesmo uma melhor calibragem dos parâmetros dos métodos em que foram usados para extrair as propriedades topográficas das redes poderiam resultar em melhores resultados.

Como trabalhos futuros pensa-se em aperfeiçoar o pré-processamento e a modelagem da rede, bem como coletar mais redações para a base de dados, a fim de aplicar um modelo preditivo sobre os dados coletados.

5. Referências

Antiqueira, Lucas, et al. "Modelando textos como redes complexas." Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana. 2005.

Cancho, R.F.; Solé, R.V. (2001) “The Small World of Human Language”, Proceedings of The Royal Society of London. Series B, Biological Sciences, 268, 2261-2265.

Sigman, M.; Cecchi, G.A. (2002) “Global Organization of the Wordnet Lexicon”, Proceedings of the National Academy of Sciences, 99, 1742-1747.

Motter, A.E.; Moura, A.P.S.; Lai, Y.C.; Dasgupta, P. (2002) “Topology of the Conceptual Network of Language”, Phys. Rev. E, 65, 065102.

Miller, G.A. (1985) “Wordnet: a dictionary browser”, Proceedings of the First International Conference on Information in Data. University of Waterloo.

Costa, L.F. (2003) “What’s in a Name?”, Int. J. Mod. Phys. C, Vol. 15, No. 1, 371-379, cond-mat/0309266.