Coh-Metrix PT-BR: Uma API web de análise textual para a educação

Raissa Camelo[1], **Samuel Justino** [1], **Rafael Ferreira Mello**[1]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)

{srtacamelo,rksamuj,rafaelflmello}@gmail.com

Resumo. *O Coh-Metrix é um sistema computacional que provê diferentes medidas de análise textual incluindo legibilidade, coerência e coesão textual. Essas medidas permitem uma análise mais profunda de diferentes tipos de textos educacionais como redações, respostas de perguntas abertas, mensagens em fóruns educacionais. Este artigo apresenta o protótipo, site e API, com a adaptação das medidas do Coh-Metrix para a língua portuguesa do Brasil.*

Abstract. *Coh-Metrix is a computational system that provides different measures of textual analysis, including legibility, coherence and textual cohesion. These measures allow a more in-depth analysis of different types of educational texts such as essays, answers to open questions, messages in educational forums. This paper describes the features of a prototype, which encompass a website and an API, of a Brazilian Portuguese version of Coh-Metrix measures.*

1. Cenário de Uso

O Coh-Metrix [1] é uma ferramenta de análise textual focada em coerência e coesão para textos escritos e falados [Graesser et al. 2004]. Amplamente utilizada na área educacional por ser uma ferramenta bastante robusta e completa, o Coh-Metrix vem ganhando destaque na sub-área de *Learning Analytics*.[Lei et al. 2014] Atualmente disponível em sua versão completa em inglês (EN), conta com 108 características distribuídas em 11 seções, cada uma extraindo valores numéricos que remetem à sintaxe, semântica, legibilidade, coesão e coerência textual. Uma versão para o português brasileiro (PT-BR) foi proposta no passado[2], porém a mesma não esta completa, contando com apenas 48 características.

O Coh-Metrix vem sendo aplicado como pré-processamento textual para diferentes projetos de tecnologia educacionais. A exemplo, trabalhos de análise automática de redações, identificação de plágios, feedback automático para cursos a distância e classificação de complexidade textual. Ao longo dos anos a procura pela ferramenta para diversos tipos de projetos de mineração de dados educacionais vem aumentando.[Dowell et al. 2016]

A análise automática de redações pode ser utilizada tanto para facilitar o trabalho de correção de uma banca ou de um professor como também servir de instrumento de auto avaliação para estudantes. O Coh-Metrix possui métricas que avaliam bem a coesão e a coerência textual, assim como a diversidade léxica de um texto fornecido,

[1]http://cohmetrix.com/
[2]http://143.107.183.175:22680/

de forma que se torna bastante eficaz utiliza-lo sobre esse contexto como apontado em [Latifi and Gierl 2020]. Não só no trabalho de [Latifi and Gierl 2020] foi utilizado o Coh-Metrix para análise de textos educacionais como também em [Wolfe et al. 2018] o Coh-Metrix foi utilizado para analisar automaticamente diálogos de mensagens trocadas entre mulheres e o *BRCA*, um sistema de tutoreamento automático que auxilia mulheres a se testarem para o câncer de mama, oferecendo informações sobre procedimentos médicos, testes de toque e o que fazer caso a mesma suspeite de um câncer. O trabalho de [Wolfe et al. 2018] foca em transmissão de informações sobre saúde e pode ser adaptado para um contexto acadêmico.

A aferição de aprendizagem através de técnicas de mineração de dados permite que professores de disciplinas EAD possam acompanhar os seus alunos de maneira mais dinâmica e eficiente. Técnicas de extração de dados de fóruns online podem detectar o nível de aprendizagem que um grupo de alunos possui em determinado assunto da matéria. O Coh-Metrix também demonstrou ser eficaz para tais tipos de analise principalmente para a detecção de presença cognitiva em fóruns educacionais [McKlin 2004][Barbosa et al. 2020]. Tal análise permite que se tenha um sumário da participação dos alunos em uma disciplina e também do índice de retenção de conteúdo pelos mesmos.

Não obstante uma versão em espanhol do Coh-Metrix com 45 características de legibilidade foi criada por pesquisadores da universidade católica do Peru [Quispesaravia et al. 2016]. Neste trabalho é ressaltada a usabilidade do Coh-Metrix para a avaliação de textos educacionais. [Quispesaravia et al. 2016] testou sua versão do Coh-Metrix classificando textos em espanhol de vários níveis escolares quanto a sua complexidade textual. O emprego do Coh-Metrix para a classificação e análise de complexidade de textos acadêmicos e escolares já foi discutido inclusive pela própria autora da ferramenta original [McCarthy et al. 2019], que realizou uma série de experimentos para aferir o nível de coesão e dificuldade de leitura de textos em livros escolares do ensino médio.

Dado o potencial do Coh-Metrix como ferramenta textual aplicável em diferentes contextos educacionais, uma versão completa para o idioma português se faz bastante conveniente. Visando adaptar o Coh-Metrix completo para a língua portuguesa do Brasil, o grupo de pesquisas da UFRPE, AiboxLab [3] inicializou o desenvolvimento de uma versão PT-BR do Coh-metrix. O projeto teve seu inicio em agosto de 2019, auxiliado pelo programa de bolsas da FACEPE [4] (PIBIC). A ideia é que o Coh-Metrix PT-BR seja disponibilizado e amplamente utilizado por pesquisadores da área, de forma a contribuir com a criação de diferentes ferramentas educacionais subsidiadas por análises textuais.

A distribuição de uma ferramenta como o Coh-Metrix para a língua portuguesa respalda novas oportunidades de analise textos educacionais. Pensando nisso, este projeto compreende não só o desenvolvimento de uma versão completa do Coh-Metrix PT-BR mas também a criação de uma API web onde a ferramenta poderá ser acessada por todos que desejarem. Este artigo se trata de uma breve amostra do protótipo de uma API WEB do Coh-Metrix PT-BR. Visando assim demonstrar como a ferramenta será oferecida em sua forma final e como a mesma deverá ser utilizada.

[3] https://aiboxlab.org/
[4] http://www.facepe.br/

2. Desenvolvimento

O Coh-Metrix PT-BR foi desenvolvido em Python. O Backend da ferramenta consiste em várias funções que executam procedimentos com o texto fornecido na entrada e retornam um valor numérico na saída. Cada função é uma característica do Coh-Metrix. Atualmente o Coh-Metrix PT-BR possui 10 seções do Coh-Metrix original. A seguir encontra-se uma breve descrição de cada seção.

1. **Descritiva**: Os índices descritivos consistem em características numéricas do texto como quantidade de palavras e parágrafos. Esses valores permitem a interpretação de padrões de dados nos textos.
2. **Coesão referencial**: Coesão referencial se trata de uma seção de características que buscam sobreposições de palavras entre orações em textos.
3. **Latent Semantic Analysis (LSA)**: Esta seção mede o grau de sobreposição semântica entre sentenças e parágrafos de um texto, classificando o texto como de alta coesão (1) ou baixa coesão (0).
4. **Diversidade Léxica**: As características desta seção calculam a variedade de palavras únicas que ocorrem no texto em relação a quantidade total de palavras no texto. Ou seja, estimam quantas palavras diferentes (que não se repetem) existem no texto. Essas métricas servem pra estipular quão coeso o texto está.
5. **Conectivos**: Esta seção contém características que indicam a quantidade de cada tipo de conectivos no texto analisado.
6. **Modelo Situacional**: Nesta seção as características referem ao nível de representação mental fornecida pelo texto. Destaca propriedades presentes na representação mental que o leitor cria ao ser inserido no contexto do texto.
7. **Complexidade Sintática**: Esta seção se trata da geração de árvores sintáticas e associação das palavras do texto em categorias de *Part-of-Speech* POS aliando-as em grupos sintáticos. A partir das árvores e grupos gerados são calculados números que estimam o valor sintático das frases e parágrafos do texto.
8. **Densidade de padrões sintáticos**: Também referente a sintaxe, esta seção calcula a frequência de padrões sintáticos como frases verbais, nominais, tipos de palavras.
9. **Informação da Palavra**: Esta seção contabiliza a incidência de cada tipo de palavras no texto: verbos, adjetivos, advérbios, pronomes, etc.
10. **Legibilidade**: Consiste no cálculo da facilidade/dificuldade de leitura e interpretação do texto. Estes cálculos são feitos utilizando várias formulas distintas, cada característica adota uma formula diferente.

As seção do Coh-Metrix contém entre 4 a 22 características cada. Cada seção foi implementada em um arquivo separado que é chamado pela a aplicação WEB. Para a implementação das características foram utilizadas as bibliotecas: NLTK, Spacy, Text Blob e Numpy. Também foi utilizada a biblioteca Lexical Diversity [5] para as características da seção de diversidade léxica.

A língua portuguesa possui uma gramática mais robusta e complexa que o inglês, de forma que algumas características do Coh-Metrix tiveram que ser repensadas para se adaptarem ao português brasileiro. Na seção de conectivos optou-se por incluir 10 características adicionais ao conjunto de 9 já existentes na ferramenta original. A língua

[5] https://pypi.org/project/lexical-diversity/

portuguesa possui mais categorias de conectivos que o inglês, de forma que essas categorias também devem ser contempladas pela versão da ferramenta PT-BR.

As 10 características adicionais da seção de conectivos são:

- CNCAlter: Incidência de conjunções alternativas
- CNCConclu: Incidência de conjunções conclusivas
- CNCExpli: Incidência de conjunções explicativas
- CNCConce: Incidência de conjunções concessiva
- CNCCondi: Incidência de conjunções condicional
- CNCConfor: Incidência de conjunções conformativas
- CNCFinal: Incidência de conjunções finais
- CNCProp: Incidência de conjunções proporcionais
- CNCComp: Incidência de conjunções comparativas
- CNCConse: Incidência de conjunções consecutivas

As características da seção de complexidade textual (Score de Facilidade de Leitura de Componentes Principais) ainda estão em desenvolvimento e necessitaram da aquisição de uma base de dados de textos educacionais para o treinamentos do modelo PCA. No artigo de McNamara (autora do Coh-Metrix original) [Graesser et al. 2011] é utilizada a base de dados TASA (*Touchstone Applied Science Associates*), para a versão em português foi escolhido o corpus de textos didáticos da USP [Murilo Gazzola 2019], uma base de dados que contém 2.076 arquivos, extraídos de exames antigos do SAEB, livros virtuais didáticos e trechos de artigos infantis de jornais. Essa base de dados foi escolhida para substituir a base TASA pois contempla as categorias de complexidades textuais propostas pela autora original do Coh-Metrix.

Cada seção do Coh-Metrix PT-BR implementada foi testada aplicando como entrada textos retirados de ambientes virtuais de aprendizagem utilizados em disciplinas da UFRPE. [Barbosa et al. 2020] Os experimentos visaram apenas aferir se as características estavam retornando valores condizentes com os textos submetidos à ferramenta.

Apesar da ferramenta ainda não contar com todas as 108 características do Coh-Metrix original, foi decido que iria-se disponibiliza-la ao público assim que possível, de forma que o desenvolvimento da ferramenta WEB teve seu inicio quando ainda se tinha apenas a metade das características funcionais. A ferramenta WEB foi sendo atualizada a medida que novas características do Coh-Metrix PT-BR eram produzidas.

Uma vez que o Coh-Metrix PT-BR foi desenvolvido em Python, foi escolhida a framework Django, da mesma língua, para adequá-lo à web. Uma API foi feita através da Django REST Framework, o que a provê uma página própria para consumo via navegador. Alternativamente, foi montada uma página de estrutura simples, funcional e de fácil interpretação através de HTML e CSS, com Django lidando com o processamento de texto para obtenção dos valores das características em funcionamento. Como há uma restrição para algumas características relacionada ao tamanho do texto, foi implementado um tratamento de exceção que além de evitar que essa restrição se torne inconveniente para o funcionamento, informa o usuário através de um resultado igual a **-1**. Essa exceção foi feita para lidar com qualquer restrição, uma vez que a ferramenta ainda passará por melhorias e podem haver erros não detectados.

No formato de API, a ferramenta retorna um JSON no qual os resultados são separados por seções e cada seção tendo suas características representadas pelo índice do

CohMetrix. Nesse formato, é usado o verbo HTTP *POST* com a chave **content** possuindo o texto submetido como valor. Já no formato de uso através do navegador não há um retorno desse tipo, pois o endpoint muda a depender da maneira que a ferramenta será utilizada.

O desenvolvimento dessa página web e da API pode ser dividido em 4 etapas, as quais são:

1. **Transformar a seção descritiva em API:** Uma vez que a seção descritiva possui características simples para realização de testes rápidos em velocidade de processamento, inicialmente foi trabalhada uma adaptação dessa seção como primeiro passo para criação da API através da Django REST Framework.
2. **Adaptar e testar as outras seções:** Tendo o primeiro passo aberto caminho para a transformação da ferramenta como um todo, essa foi realizada, criando uma API funcional com todas as seções do Coh-Metrix PT-BR, que retorna um JSON com os resultados divididos por elas e essas por suas respectivas características.
3. **Criação da página web:** Até então existia apenas a página gerada pela Django REST Framework, então foi montada a página idealizada para que se use o Coh-Metrix PT-BR de forma mais amigável para o usuário.
4. **Melhorias na ferramenta:** Para finalizar, foram feitas melhorias como uma mudança na forma de exibição dos resultados na página web, alteração de detalhes visuais, criação da página *Sobre a Ferramenta*, melhora na organização dos resultados no JSON da API e correção de erros recorrentes em algumas características.

Para ajudar no desenvolvimento, a ferramenta foi hospedada em uma máquina virtual do Google Cloud com as seguintes configurações:

Tipo	Nº de vCPUs	Memória	HD	SO
n1-highmem-8	8	52GB	10GB	Ubuntu 16.04 LTS

3. Apresentação do software

O software do Coh-Metrix PT-BR possui duas formas de uso. É possível acessar a página web, da Figura 1, na qual há uma caixa de inserção onde é posto o texto que será analisado. Após enviar o texto, os resultados são mostrados como no exemplo das Figura 3 e na Figura 4, que avisa ao usuário que se um resultado é igual a **-1**, significa que a respectiva característica não pôde ser carregada para o texto submetido. Clicando na em *Sobre a Ferramenta*, temos a página da Figura 2, cujo texto foi usado no exemplo. Nessa página, clicando em ”uso como API”se obtém o link para essa respectiva forma de uso.

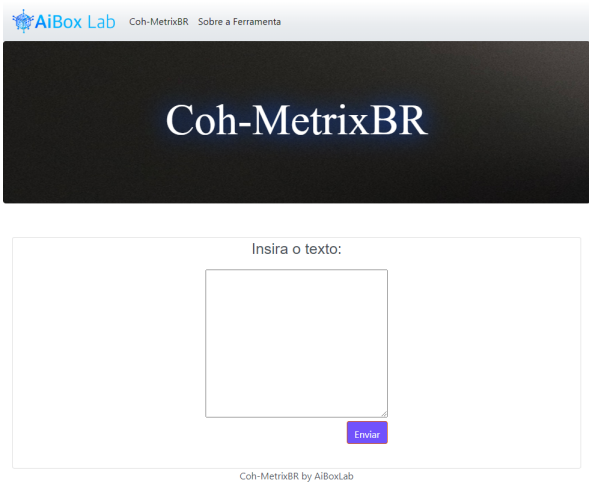

Figura 1. Página Coh-MetrixBR

Figura 2. Página Sobre a Ferramenta

4. Considerações finais

O Coh-Metrix PT-BR ainda está em fase de desenvolvimento e será disponibilizado ao público assim que sua primeira versão estiver concluída. Ao todo foram implementadas 10 das 11 seções originais do Coh-Metrix (EN). Existem muitas melhorias que podem ser feitas tanto para a otimização de desempenho da ferramenta em relação ao tempo de resposta quanto em relação a acurácia das características. A seção de Score de facilidade de texto ainda esta sendo finalizada e deverá passar por uma fase de testes antes de ser adicionada à ferramenta. Essa seção contém componentes que estipulam o grau de dificuldade/ facilidade de leitura e compreensão de um texto e é essencial para a versão final do software.

Após a construção da API foi realizado um breve teste de tempo de resposta para avaliar o desempenho do Coh-Metrix PT-BR para cada uma das 10 seções implementadas. O teste consistiu em contabilizar o tempo decorrido entre a chamada da primeira função

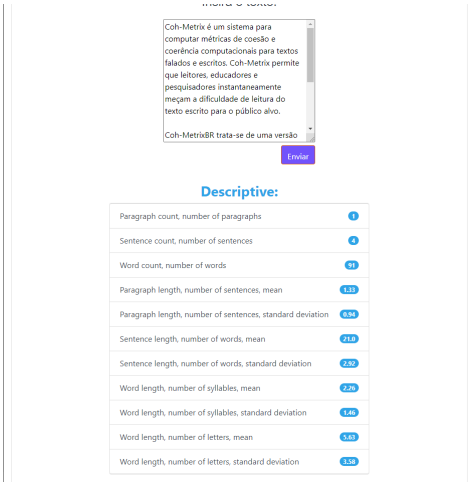

Figura 3. Topo da página de resultados

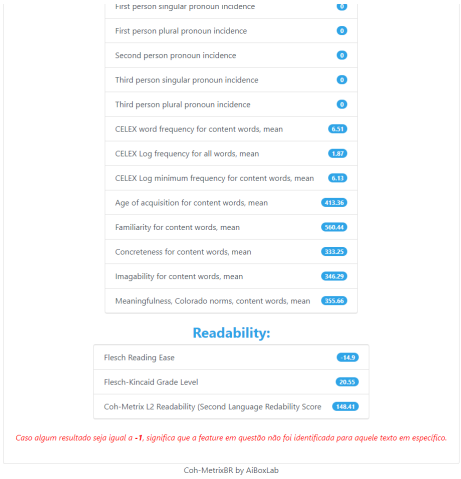

Figura 4. Base da página de resultados

(característica) de cada seção até o retorno do valor da ultima função da seção. Esse experimento foi realizado utilizando a base de dados da USP [Murilo Gazzola 2019]. Onde 100 textos aleatórios foram retirados da mesma e usados para aferir o tempo de resposta da API. A seguir pode-se observar na tabela a média, a mediana e o desvio padrão do tempo em segundos que se demanda para calcular as características de cada seção.

Seção	Média	Mediana	Desvio Padrão
Descritiva	0.041	0.040	0.008
Coesão Referencial	33.909	33.862	0.625
LSA	3.136	2.901	1.106
Diversidade Léxica	11.352	11.305	0.320
Conectivos	1.162	1.138	0.242
Modelo Situacional	0.960	0.932	0.211
Complexidade Sintática	0.924	0.898	0.205
Densidade de Padrões Sintáticos	22.566	22.473	0.508
Informação da Palavra	14.776	14.782	0.424
Legibilidade	3.423	3.536	0.386

Para as próximas etapas do projeto planeja-se não apenas finalizar a ultima seção do Coh-Metrix mas também otimizar a ferramenta, de forma que a mesma fique mais ágil e consuma menos memória para processar os dados. Muitas melhorias podem ser feitas para aprimorar a qualidade da análise e o tempo de espera por resposta da ferramenta.

A qualidade da análise textual é imprescindível para a extração de dados educacionais . Uma ferramenta como o Coh-Metrix possibilita uma visão abrangente dos componentes textuais presentes na base de dados analisada. A partir do Coh-Metrix é possível extrair informações a respeito da qualidade textual, nível de complexidade e uso de palavras especificas. Tal análise possibilita a criação de modelos cada vez mais precisos e robustos, permitindo a criação de novas ferramentas educacionais. No futuro o Coh-Metrix PT-BR será disponibilizado ao publico, permitindo que vários pesquisadores da área de mineração de textos educacionais consigam utilizar a ferramenta. Desta forma incentivando a pesquisa na área e trazendo para o português brasileiro uma ferramenta já

consolidada no estado da arte.

Referências

Barbosa, G., Camelo, R., Cavalcanti, A. P., Miranda, P., Mello, R. F., Kovanović, V., and Gašević, D. (2020). Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 605–614.

Dowell, N. M., Graesser, A. C., and Cai, Z. (2016). Language and discourse analysis with coh-metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3):72–95.

Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Latifi, S. and Gierl, M. (2020). Automated scoring of junior and senior high essays using coh-metrix features: Implications for large-scale language testing. *Language Testing*, page 0265532220929918.

Lei, C.-U., Man, K., and Ting, T. (2014). Using coh-metrix to analyse writing skills of students: A case study in a technological common core curriculum course. *Lecture Notes in Engineering and Computer Science*.

McCarthy, M., Lightman, J., Dufty, F., and McNamara, S. (2019). Using coh-metrix to assess cohesion and difficulty in high-school textbooks.

McKlin, T. E. (2004). Analyzing cognitive presence in online courses using an artificial neural network.

Murilo Gazzola, Sidney Evaldo Leal, S. M. A. (2019). Predição da complexidade textual de recursos educacionais abertos em português. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology*.

Quispesaravia, A., Perez, W., Cabezudo, M. S., and Alva-Manchego, F. (2016). Coh-metrix-esp: A complexity analysis tool for documents written in spanish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4694–4698.

Wolfe, C. R., Widmer, C. L., Torrese, C. V., and Dandignac, M. (2018). A method for automatically analyzing intelligent tutoring system dialogues with coh-metrix. *Journal of Learning Analytics*, 5(3):222–234.