Raissa Camelo Salhab

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

Recife

2019

Raissa Camelo Salhab

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

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Rafael Ferreira Leite de Mello

Recife

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

R159c Camelo , Raissa
Coh-Metrix PT-BR: Uma API web de análise textualpara à educação / Raissa Camelo . - 2021.
21 f. : il.

Orientador: Rafael Ferreira Leite de Mello.
Inclui referências e apêndice(s).

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2021.

1. Mineração de Texto. 2. CohMetrix. 3. Processamento de Linguagem Natural. 4. Inteligência Artificial. 5. Educação. I. Mello, Rafael Ferreira Leite de, orient. II. Título

CDD 004

FICHA DE APROVAÇÃO DO TRABALHO DE CONCLUSÃO DE CURSO

Trabalho defendido por Raissa Camelo Salhab às 14 horas do dia 02 de março de 2021, no link https://meet.google.com/hdh-aqmm-ens, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado *Coh-Metrix PT-BR: Uma API web de análise textual para à educação*, orientado por Rafael Ferreira Leite de Mello e aprovado pela seguinte banca examinadora:

Rafael Ferreira Leite de Mello
DC/UFRPE

André Câmara Alves do Nascimento
DC/UFRPE

À Rafael Ferreira Leite de Mello, Hertz Mariano e Daniele da Costa Camelo.

Agradecimentos

Eu acredito fortemente que todas as pessoas que passam por nossas vidas são importantes para nós, pois de alguma forma, mesmo que pequena e imperceptível essas pessoas influenciam nos caminhos que tomamos, nos pensamentos que alimentamos e na forma que nós crescemos e envelhecemos. Seja o motorista do Dois Irmãos Rui Barbosa que gentilmente esperou você subir no ônibus, ao ver que você corria para alcançá-lo; ou a tia do RU que selecionou o pedaço de carne mais suculento da bandeja, naquele dia que você visivelmente não estava bem. Todas as interações humanas possuem um significado para mim e são a prova diária e definitiva de que ninguém foi feito para viver sozinho, isolado. Somos seres sociais não só porque dependemos uns dos outros mas também porque as coisas mais magníficas que somos capazes de fazer só podem ser feitas em conjunto.

Existem pessoas cujo o nome desconheço e cuja a aparência não sou capaz de me lembrar, mas que me ajudaram a estar onde eu estou agora e a ser quem eu sou nesse exato momento, por tanto expresso aqui minha eterna gratidão por fazer parte de uma sociedade que permitiu que essas pessoas anônimas cruzassem seus caminhos com o meu, na hora certa e no local certo. Felizmente, há também um grupo enorme de pessoas cujo nome e aparência eu me lembro muito bem e será impossível esquecer. E aqui eu deposito toda minha gratidão e reconhecimento a essas pessoas. A Rural para mim foi e sempre será uma mãe, que me acolheu e me permitiu crescer. Grande parte desse sentimento se deve a essas pessoas, que irão ficar em minha vida para sempre:

Carlos Ferraz e Mª Fernanda Souza foram os primeiros amigos que fiz na rural. Carlos me ensinando uma preciosa lição, sobre não se afogar com problemas que não estão mais em suas mãos e Fernanda sempre me inspirando, sendo a pessoa maravilhosa, forte e guerreira que ela é. Caio Viana e Ricardo Luna que sempre dedicaram de seu tempo para me dar aquela força nas disciplinas e Alexandre Gouveia que me fez por muitos meses companhia e suporte na sala 42. E é claro, sempre existe espaço para novas amizades e fico muito feliz de me formar tendo Larissa Feliciana, que é uma das pessoas mais inteligentes que já conheci e que me ajudou bastante, Jen Horng, Eder Lucena e Pedro Assis como amigos. Wilber Antônio e Matheus Agostinho jamais ficarão para trás, sendo os melhores companheiros que alguém pode ter para uma aventura e para a hora do aperreio também.

Eu sempre me senti a pessoa mais sortuda do mundo, privilegiada, sempre pude correr atrás dos meus objetivos e sempre fui muito incentivada pelas pessoas ao meu redor. Dentre as pessoas que me incentivaram e que também são responsáveis por manter

a minha chama acesa (vulgo sangue nos olhos) estão: Suzana Sampaio e Jeane Melo que desde sempre acreditaram em mim e no meu potencial, André Aziz, Abner Barros, Péricles Miranda e Valmir Macário, que sempre estiveram dispostos a me ajudar, se mostrando professores excepcionais. Sandra Xavier, sempre presente durante toda minha jornada em BCC, sendo a rainha das secretárias e providências. Luciano Pacifico, pelas preciosas dicas de vida e convivência com terceiros. E é claro a Filipe Rolim, que não só me acolheu como orientanda nos meus primeiros anos de faculdade mas que também esteve presente numa das fases mais importantes da minha vida.
Por fim, às três pessoas mais importantes para a conclusão deste trabalho: Rafael Ferreira, que foi meu orientador nesses últimos 2 anos, abrindo meus olhos para um mundo de possibilidades e para a real natureza da ciência, a colaboração. Hertz Mariano que foi um tutor e é um grande amigo, que me incentivou a buscar uma carreira acadêmica e a quem eu devo a alcunha “Senhorita Camelo”, que uso até hoje. E é claro, um agradecimento especial para minha heroína favorita, minha mãe, Daniele da Costa Camelo. Sem vocês esse trabalho não seria possível.

“A persistência é o caminho do êxito.”
(Charles Chaplin)

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 e 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.

Palavras-Chave: CohMetrix, Mineração de Texto, Processamento de Linguagem Natural.

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 and 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.

Keywords: CohMetrix, Text Mining, Natural Language Processing.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

PT	Português
BR	Brasileiro
EN	Inglês
BRCA	*BReast CAncer and Genetics Intelligent Semantic Tutoring*
EAD	Educação à Distância
FACEPE	Fundação do Amparo a Ciência e Tecnologia
PIBIC	Programa Institucional de Bolsas de Iniciação Científica
API	*Application Programming Interface*
LSA	*Latent Semantic Analysis*
NLTK	*Natural language Toolkit*
SAEB	Sistema de Avaliação da Educação Básica
TASA	*Touchstone Applied Science Associates*

Sumário

1 Introdução

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; MAN; TING, 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 está completa, contando com apenas 48 características.

O Coh-Metrix vem sendo aplicado como pré-processamento textual para diferentes projetos de tecnologia educacionais. Por 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; GRAESSER; CAI, 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, de forma que se torna bastante eficaz utiliza-lo sobre esse contexto como apontado em (LATIFI; GIERL, 2020). Não só no trabalho de (LATIFI; 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 Gist *(BReast CAncer and Genetics Intelligent Semantic Tutoring)*, 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

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

assunto da matéria. O Coh-Metrix também demonstrou ser eficaz para tais tipos de análise 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 (QUISPE-SARAVIA 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 análise 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.

Para mais detalhes mais avançados a respeito das características, forma que foram implementadas e formulas utilizadas em cada sessão, o livro do CohMetrix original deve ser consultado (MCNAMARA et al., 2014).

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 [1] 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 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. São conjunções que indicam sentido de alternância ou exclusão: ou, ora, ora quer etc.
- CNCConclu: Incidência de conjunções conclusivas. São conjunções que indicam sentido de conclusão: logo, assim, por tanto, por isso etc.
- CNCExpli: Incidência de conjunções explicativas. São conjunções que indicam sentido de explicação: porque, que, pois (antes do verbo), porquanto etc.
- CNCConce: Incidência de conjunções concessiva. São conjunções que indicam sentido de concessão: embora, conquanto, ainda que etc.
- CNCCondi: Incidência de conjunções condicional. São conjunções que indicam sentido de condição: se, caso, conquanto que, desde que etc.
- CNCConfor: Incidência de conjunções conformativas. São conjunções que indicam sentido de conformidade: conforme, segundo, como etc.
- CNCFinal: Incidência de conjunções finais. São conjunções que indicam sentido de finalidade: para que, afim de que etc.

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

- CNCProp: Incidência de conjunções proporcionais São conjunções que indicam sentido de proporção: á medida que, ao passo que, quanto mais, quanto menos etc.
- CNCComp: Incidência de conjunções comparativas. São conjunções que indicam sentido de comparação: maior que, menor que, como, mais etc.
- CNCConse: Incidência de conjunções consecutivas. São conjunções que indicam sentido de consequência: de sorte que, de forma que, tal que, tamanho que etc.

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; MCNAMARA; KULIKOWICH, 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 (GAZZOLA SIDNEY EVALDO LEAL, 2019), uma base de dados que contém 2.076 arquivos, extraídos de exames antigos do SAEB (Sistema de Avaliação da Educação Básica), 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 a mesma seria disponibilizada 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 linguagem, para adequá-lo à web. Uma API foi feita através da Django REST Framework, oferecendo 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

Tabela 1 – Configurações do Ambeinte Web

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 3a e na Figura 3b, 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.

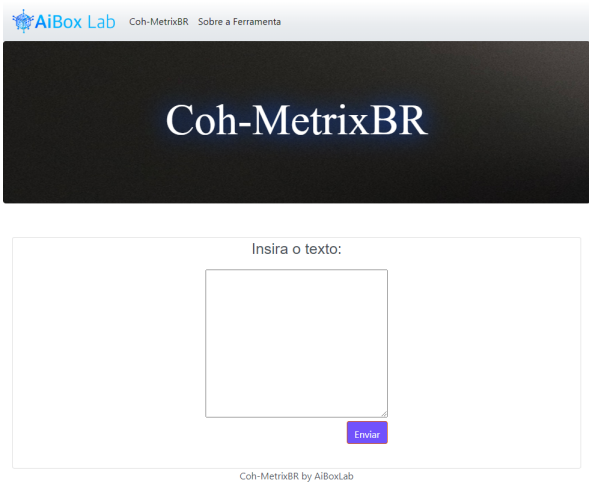
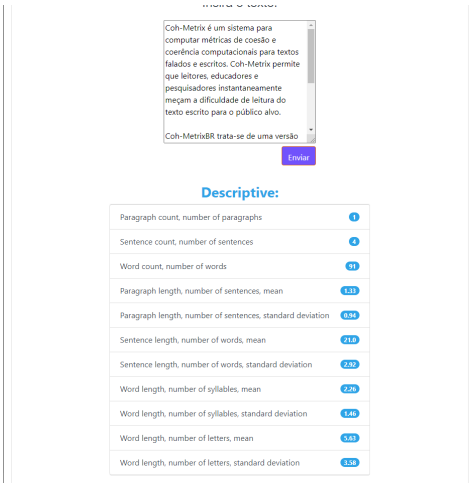
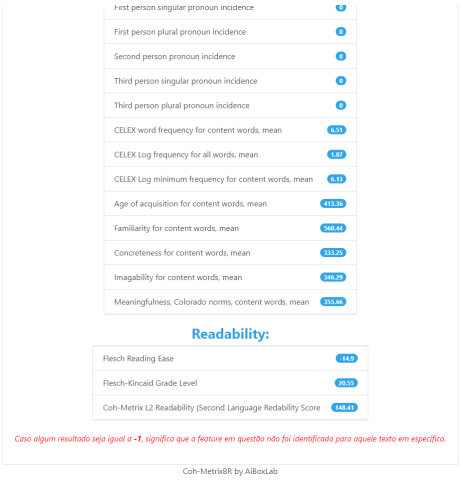


Figura 1 – Página Coh-MetrixBR

Figura 2 – Página Sobre a Ferramenta



(a) Topo da página de resultados



(b) Base da página de resultados

4 Conclusões

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 (Complexidade Textual), referenciada na seção de desenvolvimento, 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 (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 (GAZZOLA SIDNEY EVALDO LEAL, 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

Tabela 2 – Experimento CohMetrix: Tempo de resposta de cada seção (em segundos)

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 po-

dem 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. et al. Towards automatic cross-language classification of cognitive presence in online discussions. In: *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. [S.l.: s.n.], 2020. p. 605–614. Citado 2 vezes nas páginas 12 e 15.

DOWELL, N. M.; GRAESSER, A. C.; CAI, Z. Language and discourse analysis with coh-metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, v. 3, n. 3, p. 72–95, 2016. Citado na página 11.

GAZZOLA SIDNEY EVALDO LEAL, S. M. A. M. Predição da complexidade textual de recursos educacionais abertos em português. In: *Proceedings of the Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 15 e 19.

GRAESSER, A. C.; MCNAMARA, D. S.; KULIKOWICH, J. M. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, Sage Publications Sage CA: Los Angeles, CA, v. 40, n. 5, p. 223–234, 2011. Citado na página 15.

GRAESSER, A. C. et al. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, Springer, v. 36, n. 2, p. 193–202, 2004. Citado na página 11.

LATIFI, S.; GIERL, M. Automated scoring of junior and senior high essays using coh-metrix features: Implications for large-scale language testing. *Language Testing*, SAGE Publications Sage UK: London, England, p. 0265532220929918, 2020. Citado na página 11.

LEI, C.-U.; MAN, K.; TING, T. 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*, Newswood Ltd.., 2014. Citado na página 11.

MCCARTHY, M. et al. Using coh-metrix to assess cohesion and difficulty in high-school textbooks. Citeseer, 2019. Citado na página 12.

MCKLIN, T. E. Analyzing cognitive presence in online courses using an artificial neural network. 2004. Citado na página 12.

MCNAMARA, D. S. et al. *Automated evaluation of text and discourse with Coh-Metrix*. [S.l.]: Cambridge University Press, 2014. Citado na página 14.

QUISPESARAVIA, A. et al. 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)*. [S.l.: s.n.], 2016. p. 4694–4698. Citado na página 12.

WOLFE, C. R. et al. A method for automatically analyzing intelligent tutoring system dialogues with coh-metrix. *Journal of Learning Analytics*, ERIC, v. 5, n. 3, p. 222–234, 2018. Citado na página 11.