UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE CIÊNCIAS MATEMÁTICAS
E DE COMPUTAÇÃO

SCC5908 INTRODUÇÃO AO PROCESSAMENTO DE LÍNGUA NATURAL
&
SCC0633 PROCESSAMENTO DE LINGUAGEM NATURAL

Etapa 2 do projeto - Implementação

Alunos:
8436630 - Murilo Gleyson Gazzola
8936801 - Otávio Augusto Ferreira Sousa
9293032- Raul Wagner Martins Costa
8628220 - Vitor Albuquerque de Paula

Professores:
Prof. Dr. Thiago A. S. Pardo
Profa. Dra. Maria das Graças V. Nunes

USP – São Carlos
Junho de 2018

Sumário

1 Introdução

Existem grandes acervos on-line que não se encontram categorizados, como, por exemplo, a *Wikilivros* [1] (Figura 1). Um classificador automático tem como finalidade auxiliar na categorização dos materiais, seja por nível de dificuldade, assunto, linguagem, entre varias outros de forma a melhorar a organização e tornar o conteúdo mais acessível.

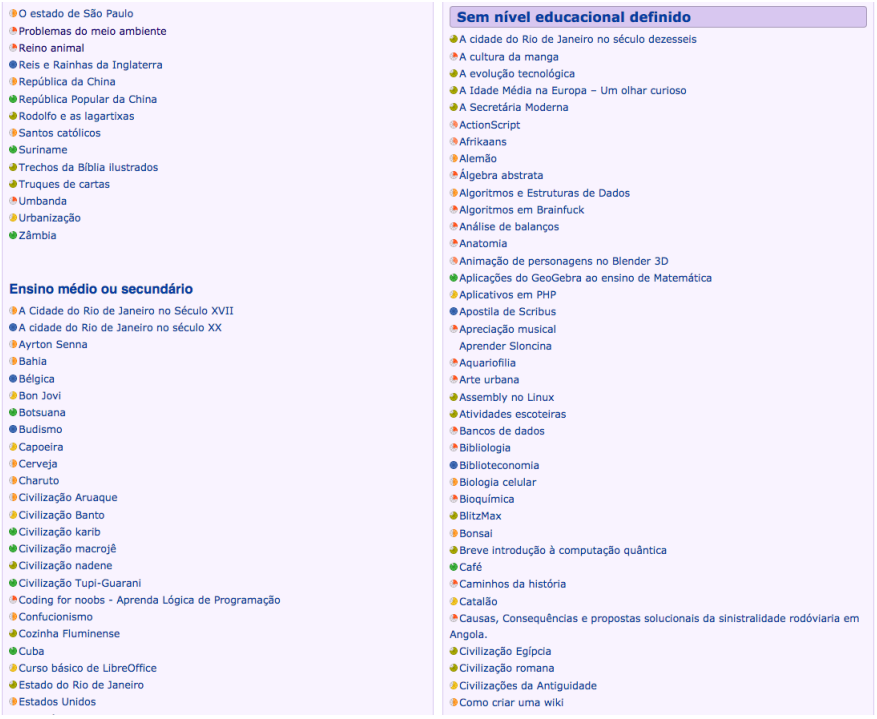


Figura 1: Wikilivros em Português - sem nível escolar

A proposta deste projeto é fazer a classificação automática do nível de textos escolares, predizendo se os textos são adequados ao ensino fundamental ou ensino médio. Nessa etapa, foi criado um preditor de nível escolar, viabilizando tais motivações.

2 Objetivo

Este projeto visa classificar documentos em relação aos seus níveis de escolaridade através do uso de aprendizado de máquina. Em particular, as técnicas de aprendizado supervisionado *Support Vector Machine* (SVM) e *Multilayer Perceptron* (MLP) são utilizadas para classificar as instâncias de documentos entre nível de ensino fundamental e nível de ensino médio.

3 Extração do córpus

O corpus foi extraído a partir dos livros didáticos digitalizados em *Portable Document Format* (PDF) do site do Ministério da Educação (MEC) [2].

[1] `https://pt.wikibooks.org/wiki/Wikilivros`
[2] Disponível em `https://goo.gl/yNtJP7`

Para extração do texto, foram utilizadas ferramentas de conversão *PDF* para *txt*. Contudo, esse método gera falhas no texto, fazendo-se necessário o uso de Expressões Regulares (ER) e a correção manual.

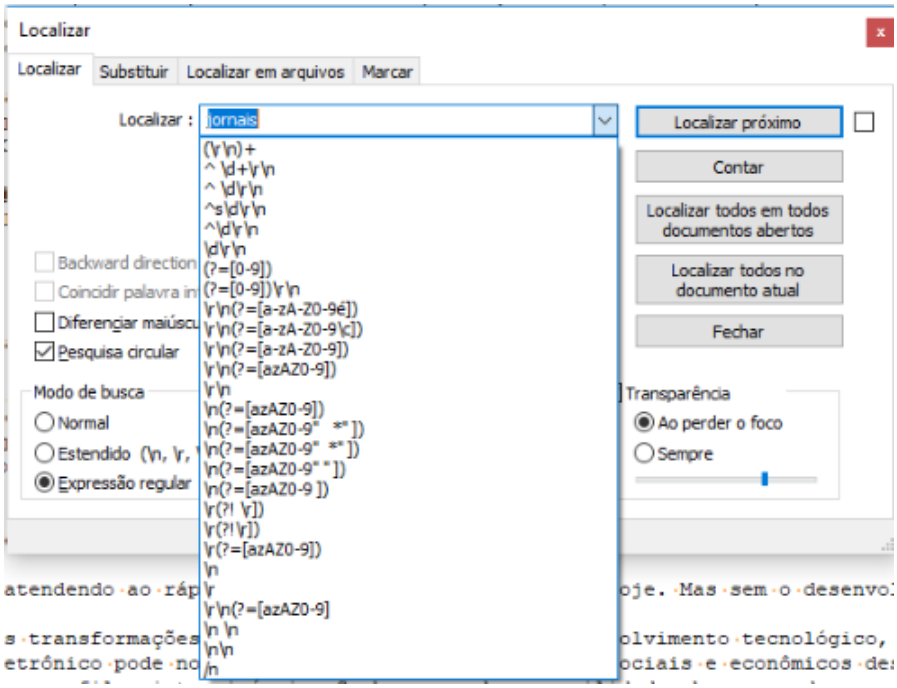


Figura 2: Exemplos de expressões ER utilizadas no *notepad++*

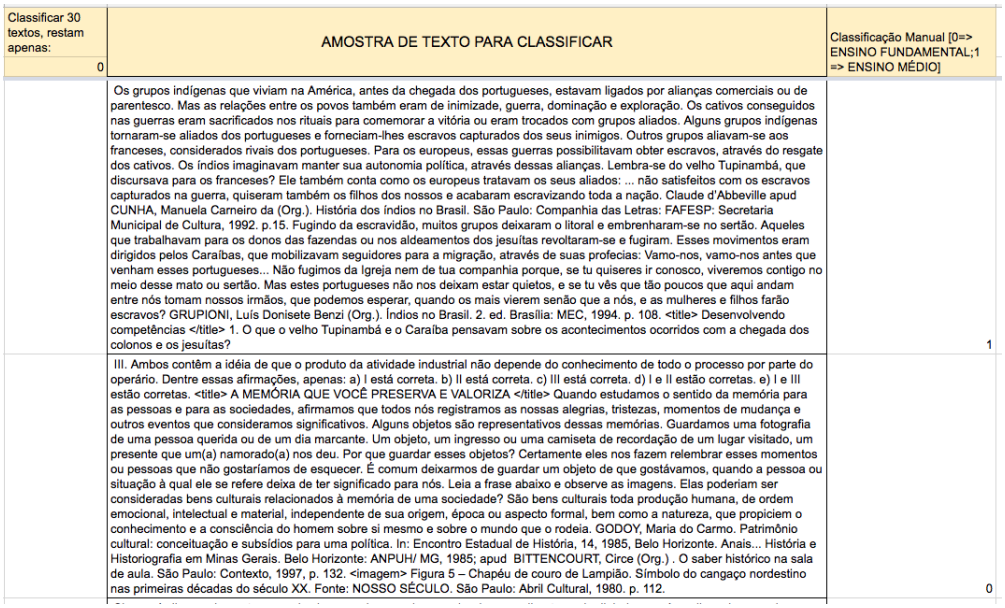
O córpus criado possui 125.096 palavras, separadas em amostras com no mínimo 300 palavras, máximo 575 palavras e média de 337 palavras.

O córpus para o Ensino Fundamental possui 57.385 palavras e para o Ensino Médio 67.711 palavras. Também, todo o corpus possui 7.748 sentenças.

4 Anotação do córpus

Como o foco deste trabalho é realizar a classificação de documentos, a anotação do córpus consiste na geração dos rótulos para cada amostra criada na fase de extração. No entanto, a anotação não foi feita manualmente, pois utilizamos os rótulos definidos pelo Ministério da Educação (MEC). Assim, para que pudéssemos estudar o processo de anotação de um córpus, foi proposto que algumas amostras fossem selecionadas e, posteriormente, rotuladas pelos anotadores de forma manual.

A anotação manual do córpus foi realizada por quatro pessoas. A fim de evitar viés nessa fase, foram selecionadas ao todo trinta amostras aleatoriamente entre as amostras dos ensinos fundamental e médio. Em seguida, os anotadores rotularam as amostras baseando-se apenas no conhecimento de mundo de cada um, ou seja, não seguiram métrica ou algoritmo algum. Quanto à interface de anotação, utilizamos a plataforma compartilhada Google Sheets e organizamos as planilhas adequadamente para a tarefa. Veja na Figura 3, a interface utilizada para a anotação manual.



Classificar 30 textos, restam apenas: 0	AMOSTRA DE TEXTO PARA CLASSIFICAR	Classificação Manual [0=> ENSINO FUNDAMENTAL;1 => ENSINO MÉDIO]
	Os grupos indígenas que viviam na América, antes da chegada dos portugueses, estavam ligados por alianças comerciais ou de parentesco. Mas as relações entre os povos também eram de inimizade, guerra, dominação e exploração. Os cativos conseguidos nas guerras eram sacrificados nos rituais para comemorar a vitória ou eram trocados com grupos aliados. Alguns grupos indígenas tornaram-se aliados dos portugueses e forneciam-lhes escravos capturados dos seus inimigos. Outros grupos aliavam-se aos franceses, considerados rivais dos portugueses. Para os europeus, essas guerras possibilitavam obter escravos, através do resgate dos cativos. Os índios imaginavam manter sua autonomia política, através dessas alianças. Lembra-se do velho Tupinambá, que discursava para os franceses? Ele também conta como os europeus tratavam os seus aliados: ... não satisfeitos com os escravos capturados na guerra, quiseram também os filhos dos nossos e acabaram escravizando toda a nação. Claude d'Abbeville apud CUNHA, Manuela Carneiro da (Org.). História dos índios no Brasil. São Paulo: Companhia das Letras: FAFESP: Secretaria Municipal de Cultura, 1992. p.15. Fugindo da escravidão, muitos grupos deixaram o litoral e embrenharam-se no sertão. Aqueles que trabalhavam para os donos das fazendas ou nos aldeamentos dos jesuítas revoltaram-se e fugiram. Esses movimentos eram dirigidos pelos Caraíbas, que mobilizavam seguidores para a migração, através de suas profecias: Vamo-nos, vamo-nos antes que venham esses portugueses... Não fugimos da Igreja nem de tua companhia porque, se tu quiseres ir conosco, viveremos contigo no meio desse mato ou sertão. Mas estes portugueses não nos deixam estar quietos, e se tu vês que tão poucos que aqui andam entre nós tomam nossos irmãos, que podemos esperar, quando os mais vierem senão que a nós, e as mulheres e filhos farão escravos? GRUPIONI, Luís Donisete Benzi (Org.). Índios no Brasil. 2. ed. Brasília: MEC, 1994. p. 108. <title> Desenvolvendo competências </title> 1. O que o velho Tupinambá e o Caraíba pensavam sobre os acontecimentos ocorridos com a chegada dos colonos e os jesuítas?	1
	III. Ambos contêm a idéia de que o produto da atividade industrial não depende do conhecimento de todo o processo por parte do operário. Dentre essas afirmações, apenas: a) I está correta. b) II está correta. c) III está correta. d) I e II estão corretas. e) I e III estão corretas. <title> A MEMÓRIA QUE VOCÊ PRESERVA E VALORIZA </title> Quando estudamos o sentido da memória para as pessoas e para as sociedades, afirmamos que todos nós registramos as nossas alegrias, tristezas, momentos de mudança e outros eventos que consideramos significativos. Alguns objetos são representativos dessas memórias. Guardamos uma fotografia de uma pessoa querida ou de um dia marcante. Um objeto, um ingresso ou uma camiseta de recordação de um lugar visitado, um presente que um(a) namorado(a) nos deu. Por que guardar esses objetos? Certamente eles nos fazem relembrar esses momentos ou pessoas que não gostaríamos de esquecer. É comum deixarmos de guardar um objeto de que gostávamos, quando a pessoa ou situação à qual ele se refere deixa de ter significado para nós. Leia a frase abaixo e observe as imagens. Elas poderiam ser consideradas bens culturais relacionados à memória de uma sociedade? São bens culturais toda produção humana, de ordem emocional, intelectual e material, independente de sua origem, época ou aspecto formal, bem como a natureza, que propiciem o conhecimento e a consciência do homem sobre si mesmo e sobre o mundo que o rodeia. GODOY, Maria do Carmo. Patrimônio cultural: conceituação e subsídios para uma política. In: Encontro Estadual de História, 14, 1985, Belo Horizonte. Anais... História e Historiografia em Minas Gerais. Belo Horizonte: ANPUH/ MG, 1985; apud BITTENCOURT, Circe (Org.) . O saber histórico na sala de aula. São Paulo: Contexto, 1997, p. 132. <imagem> Figura 5 – Chapéu de couro de Lampião. Símbolo do cangaço nordestino nas primeiras décadas do século XX. Fonte: NOSSO SÉCULO. São Paulo: Abril Cultural, 1980. p. 112.	0

Figura 3: Interface de anotação

A medida de concordância entre os anotadores foi obtida através do coeficiente Kappa de Cohen. Os valores de concordância para cada par de anotadores são: 0.200, 0.067, 0.133, 0.139, 0.217 e 0.286. Para obter a concordância entre todos anotadores, foi feita a média aritmética entre os coeficientes Kappa calculados para cada par de anotadores. Dessa forma, a concordância medida entre todos anotadores é igual a 0.174. Apenas em 26.7% das amostras houve concordância total entre os quatro anotadores.

Além disso, vale dizer que a melhor taxa de acerto dos anotadores, individualmente, foi menor que 60%. Com base nos resultados de baixa concordância e taxa de acerto moderada, podemos observar o quão difícil é essa tarefa de classificação.

5 Desenvolvimento

5.1 Atividades Realizadas

Neste trabalho foi criado um classificador automático para predizer o nível escolar de um material educacional. Foram selecionados apenas 2 níveis escolares (1 - Ensino Fundamental e 2 - Ensino Médio) e foram restringidos os domínios das amostras de textos do conhecimento escolar da seguinte forma: História e Geografia para Ensino Fundamental e Ciências Humanas para o Ensino Médio.

Em seguida, foram selecionados 64 métricas mostradas nas tabelas de 1 a 12 obtidas do Coh-Metrix-Port [1] e do Coh-Metrix-Dementia [2]. A relação de todas as métricas está na Apêndice A. As métricas de disfluências do Coh-Metrix-Dementia não foram usadas, pois não há necessidade para este tipo de tarefa.

Além disso, os grupos de métricas foram agrupadas em 12 categorias dividas em **Contagens Básicas** (A.1 até A.14), **Operadores Lógicos** (A.15 até A.19), **Frequência de Palavras de Conteúdo** (A.20 até A.21), **Hiperônimos** (A.22), **Pronomes e *Types*** (A.23 até A.28), **Conectivos** (A.31 até A.39), **Ambiguidade** (A.40 até A.46), **Densidade Semântica** (A.64), **Constituintes** (A.29 até A.30), **Anáforas** (A.52 até A.53), **Correferências** (A.47 até A.51) e **Análise da Semântica Latente** (A.54 até A.63).

Tabela 1: Contagens Básicas

Incidência de Adjetivos	Incidência de Advérbios	Incidência de Palavras de Conteúdo
Incidência de Adjetivos	Incidência de Advérbios	Incidência de Palavras de Conteúdo
Índice Flesch	Incidência de Palavras Funcionais	Sentenças por Parágrafos
Sílabas por Palavra de Conteúdo	Palavras por Sentenças	Incidência de Substantivos
Número de Parágrafos	Número de Sentenças	Número de Palavras
Incidência de Pronomes	Incidência de Verbos	

Tabela 2: Operadores Lógicos

Incidência do operador lógico E	Incidência do operador lógico OU	Incidência de Operadores Lógicos
Incidência do operador lógico SE	Incidência de negação	

Tabela 3: Frequência de Palavras de Conteúdo

Frequência das palavras de conteúdo	Frequência da palavra de conteúdo mais rara

Tabela 4: Hiperônimos.

Hiperônimos de Verbos

Tabela 5: Pronomes e *Types*

Índice de Brunet	Estatística de Horoné	Cláusulas por Sentença
Incidência de Pronomes Pessoais	Pronomes por Sintagmas	Relação Tipo por Token (Type)

Tabela 6: Conectivos

Incidência de Conectivos	Incidência de conectivos classificados como aditivos negativos	Incidência de conectivos classificados como aditivos positivos
Incidência de conectivos classificados como causais negativos	Incidência de conectivos classificados como causais positivos	Conectivos Lógicos Negativos
Conectivos Lógicos Positivos	Incidência de conectivos classificados como temporais negativos	Incidência de conectivos classificados como temporais positivos

Tabela 7: Ambiguidade

Ambiguidade de Adjetivos	Ambiguidade de Advérbios	Ambiguidade de Substantivos
Ambiguidade de Verbos	Distância de Dependência	Complexidade de Frazier
Complexidade de Yngve		

Tabela 8: Densidade Semântica.

Densidade de Conteúdo

Tabela 9: Constituintes

Incidência de Sintagmas (Constituintes)	Palavras Antes de Verbos Principais (Constituintes)

Tabela 10: Anáforas

Referência Anafórica Adjacente	Referência Anafórica

Tabela 11: Correferências

Sobreposição de Argumentos Adjacentes	Sobreposição de Argumentos	Sobreposição de Radicais de palavras Adjacentes
Sobreposição de Radicais de palavras	Sobreposição de Palavras de conteúdo em sentenças adjacentes	

Tabela 12: Análise da Semântica Latente

LSA: média entre sentenças adjacentes	LSA: desvio padrão entre sentenças adjacentes
LSA: média entre sentenças todos os pares de sentenças	LSA: desvio padrão entre sentenças todos os pares de sentenças
LSA: média entre parágrafos adjacentes	LSA: desvio padrão entre parágrafos adjacentes
LSA: média de givenness das sentenças	LSA: desvio padrão de givenness das sentenças
LSA: média do span das sentenças	LSA: desvio padrão do span das sentenças

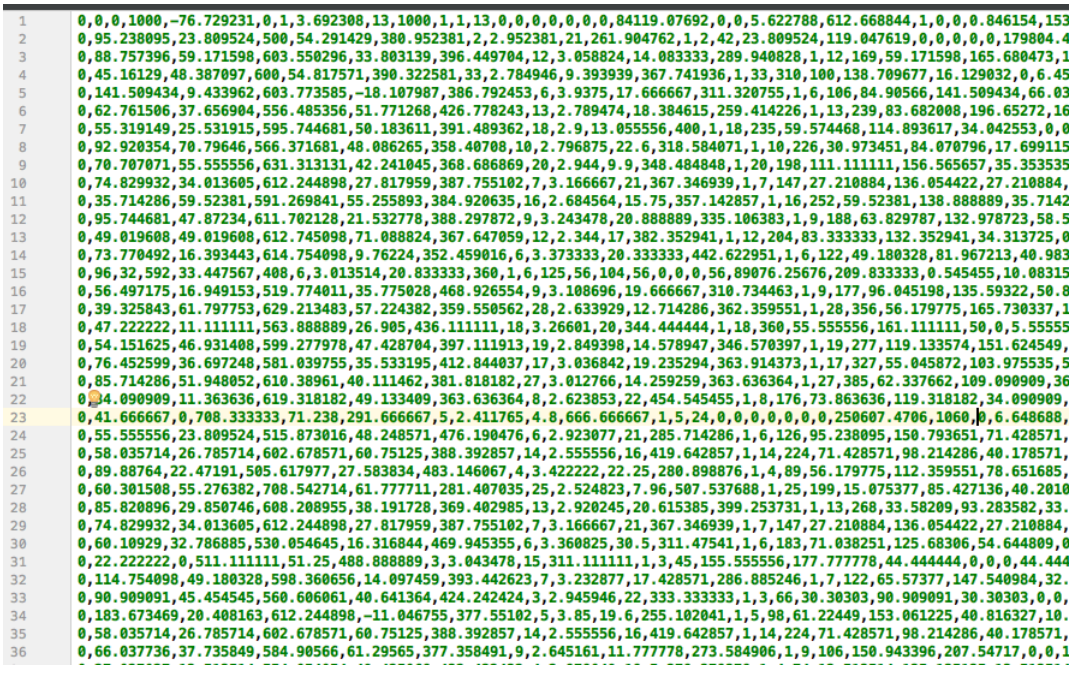


Figura 4: Extração de *features*

Após isso, foi realizado a extração de todas as métricas (Figura 4) e criado uma entrada para os classificadores utilizados.

O corpus no total possui 367 amostras (dividas em 170 para o Ensino Fundamental e 197 para o Ensino Médio). Todas as amostras estavam classificadas. Então, para nosso grupo de treinamento utilizamos 80% para treinamento e 20% para teste. Considerando que as classes estavam balanceadas realizamos a escolha aleatória, de quais instâncias iriam para teste e quais iriam para treinamento usando o pacote SkLearn[3] (from sklearn.model_selection import train_test_split) e usando a variável semente como 2018 (*random_state*).

Foram usados dois classificadores, Máquina de Vetores de Suporte e Perceptron Multicamadas (MLP, do inglês: Multi Layer Perceptron) para os dois classificadores foram usados o modelo 20% de teste e 80% de treino. Por fim, para o melhor classificador foi usado uma validação cruzada. Os resultados da classificação usando SVM (Tabela 13) obteve o valor médio de 64,2% e representando o maior valor médio de precisão entre os classificadores do ambiente de experimento. Em relação a MLP (Tabela 14) obteve o segundo melhor resultado com a média de 60,2% de precisão. Com estes resultados, foi realizado uma validação cruzada (*cross-validation*) usando *k-fold* para avaliar a capacidade de generalização do modelo criado a partir do conjunto de dados onde o SVM obteve 68,6% de precisão, o k utilizado foi 10. Ou seja, foi divido o conjunto total de amostras em 10 subconjuntos mutualmente exclusivos do mesmo tamanho e utilizado para teste e os k-1 restantes utilizados para estimação da acurácia do modelo.

Tabela 13: Classificação usando SVM

Classe	Precisão	Recall	F-Measure
Ensino Fundamental	0,559	0,744	0,639
Ensino Médio	0,709	0,516	0,597
Média	**0,642**	0,619	0,616

Tabela 14: Classificação usando MLP

Classe	Precisão	Recall	F-Measure
Ensino Fundamental	0,536	0,669	0,595
Ensino Médio	0,656	0,522	0,581
Média	**0,602**	0,588	0,588

Tabela 15: Classificação usando SVM (*Cross Validation*)

Classe	Precisão	Recall	F-Measure
Ensino Fundamental	0,671	0,635	0,653
Ensino Médio	0,699	0,731	0,715
Média	**0,686**	0,687	0,686

5.2 Predição com exemplo da Wikilivros

Em seguida foi criado um preditor, objetivo deste trabalho, onde é realizado a entrada de uma amostra e o preditor é capaz de dizer se a amostra é do Ensino Fundamental ou

[3] `http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

Ensino Médio. Para o seguinte exemplo retirado da Wikilivros [3] categorizado no Ensino Fundamental ou Básico pela Wikilivros. Esse exemplo não está na base de treinamento e, também, não está na base de teste.

A ação do homem no meio ambiente
A ação antrópica tem se provado fundamental na modificação das condições terrestres, principalmente quanto ao desmatamento e ao ampliamento do efeito estufa. No entanto é preciso que fique claro que nem todos os desastres ambientais são causados pelo homem: abalos sísmicos e vulcanismos, por exemplo, não são influenciados por tais ações, pois ocorrerem abaixo da superfície terrestre, por isso não deve haver generalização.
Impacto ambiental
A ação do homem geralmente tem impactos ambientais negativos, porque as consequências ambientais de suas atividades são geralmente desconsideradas ou ignoradas. Além disso, acidentes que causam grandes danos ambientais podem ocorrer, como vazamento de óleo.

Amostra do Livro **"Problemas do meio ambiente"** [3]

Na Figura 5 é mostrado o resultado do preditor criado, para a amostra de livro foram extraídas as *features* e, utilizando o modelo pré-treinado com SVM (*kernel* linear) ele predisse a amostra com possibilidade de ser para o Ensino Fundamental. Também, muitas *features* chegaram próximo a zero ou zero, como é o caso das *features* de Análise Semântica Latente (LSA) e os incidentes de conectivos lógicos positivos ou negativos.

5.3 Sugestão de *Feature* Adicional

Neste trabalho, foi proposto o uso de uma *feature* adicional além das mencionadas anteriormente. Está nova *feature* consiste em calcular a média aritmética entre as profundidades das árvores sintáticas de cada sentença em um texto dado.

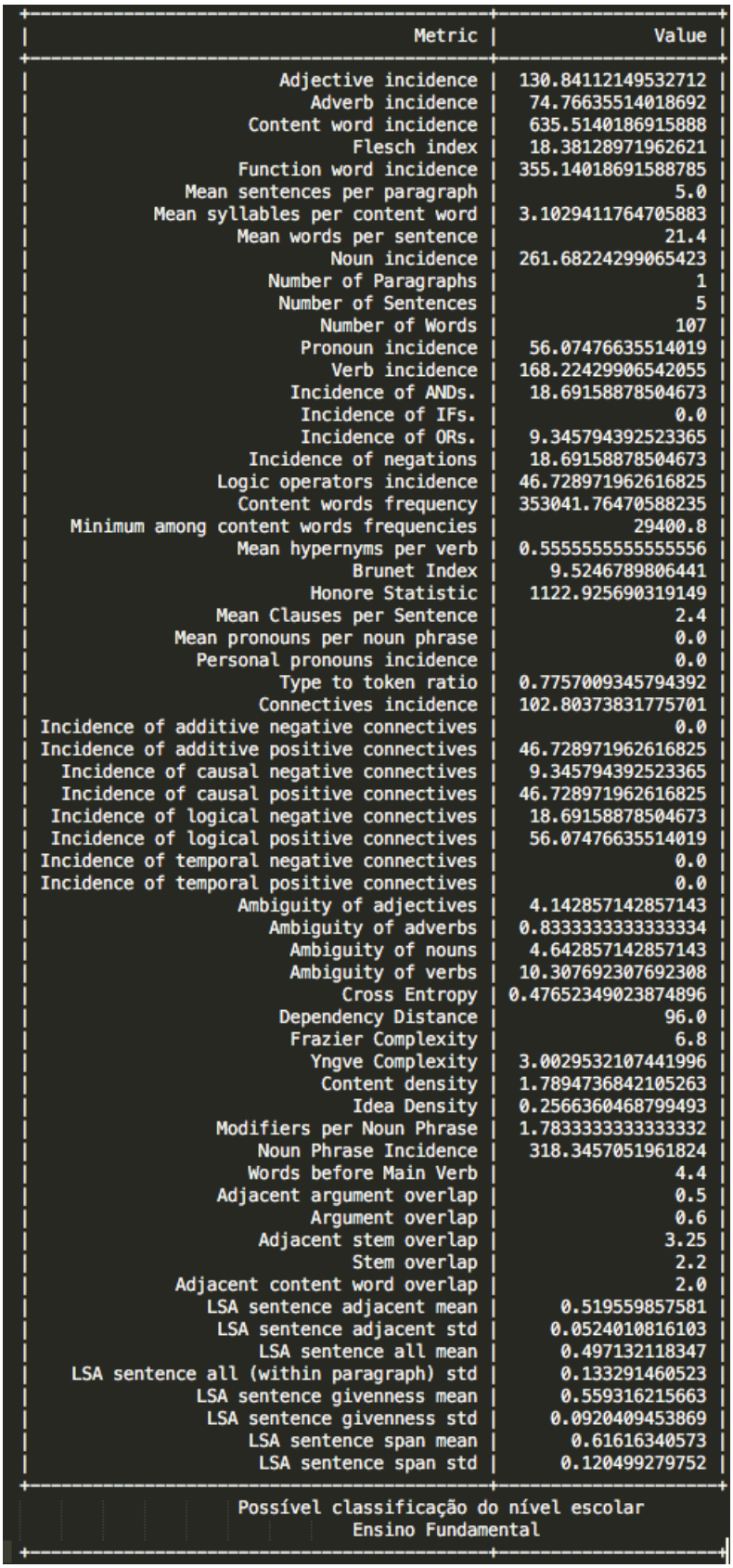
De uma forma mais clara, a *feature* foi implementada como segue. Primeiramente, um texto é recebido e dividido em sentenças. Em seguida, a árvore sintática e a profundidade da árvore são obtidas para cada sentença. Para tal, utilizamos um *parser* disponível no *site*[4] do projeto *Visual Interactive Syntax Learning* (VISL) do *Institute of Language and Communication* da *University of Southern Denmark*. Dado o resultado da análise sintática de uma sentença em forma de árvore sintática, a profundidade da árvore é obtida com o auxílio de expressão regular.

Posteriormente, consideramos que o resultado desta nova *feature* para um determinado texto equivale a média aritmética entre as profundidades das árvores sintáticas de cada sentença presente no texto. A lógica é que em textos mais fáceis, como textos do ensino fundamental, a análise sintática é mais simples, resultando em menores árvores sintáticas. Analogamente, em textos mais difíceis, a análise sintática é mais complexa e isso é refletido na profundidade das árvores sintáticas.

Tabela 16: Classificação usando SVM com nova *feature*

Classe	Precisão	Recall	F-Measure
Ensino Fundamental	0,691	0,854	0,764
Ensino Médio	0,724	0,500	0,591
Média	**0,705**	0,701	0,698

[4]visl.sdu.dk/visl/pt/parsing/automatic/trees.php



Metric	Value
Adjective incidence	130.84112149532712
Adverb incidence	74.76635514018692
Content word incidence	635.5140186915888
Flesch index	18.38128971962621
Function word incidence	355.14018691588785
Mean sentences per paragraph	5.0
Mean syllables per content word	3.1029411764705883
Mean words per sentence	21.4
Noun incidence	261.68224299065423
Number of Paragraphs	1
Number of Sentences	5
Number of Words	107
Pronoun incidence	56.07476635514019
Verb incidence	168.22429906542055
Incidence of ANDs.	18.69158878504673
Incidence of IFs.	0.0
Incidence of ORs.	9.345794392523365
Incidence of negations	18.69158878504673
Logic operators incidence	46.728971962616825
Content words frequency	353041.76470588235
Minimum among content words frequencies	29400.8
Mean hypernyms per verb	0.5555555555555556
Brunet Index	9.5246789806441
Honore Statistic	1122.925690319149
Mean Clauses per Sentence	2.4
Mean pronouns per noun phrase	0.0
Personal pronouns incidence	0.0
Type to token ratio	0.7757009345794392
Connectives incidence	102.80373831775701
Incidence of additive negative connectives	0.0
Incidence of additive positive connectives	46.728971962616825
Incidence of causal negative connectives	9.345794392523365
Incidence of causal positive connectives	46.728971962616825
Incidence of logical negative connectives	18.69158878504673
Incidence of logical positive connectives	56.07476635514019
Incidence of temporal negative connectives	0.0
Incidence of temporal positive connectives	0.0
Ambiguity of adjectives	4.142857142857143
Ambiguity of adverbs	0.8333333333333334
Ambiguity of nouns	4.642857142857143
Ambiguity of verbs	10.307692307692308
Cross Entropy	0.47652349023874896
Dependency Distance	96.0
Frazier Complexity	6.8
Yngve Complexity	3.0029532107441996
Content density	1.7894736842105263
Idea Density	0.2566360468799493
Modifiers per Noun Phrase	1.7833333333333332
Noun Phrase Incidence	318.3457051961824
Words before Main Verb	4.4
Adjacent argument overlap	0.5
Argument overlap	0.6
Adjacent stem overlap	3.25
Stem overlap	2.2
Adjacent content word overlap	2.0
LSA sentence adjacent mean	0.519559857581
LSA sentence adjacent std	0.0524010816103
LSA sentence all mean	0.497132118347
LSA sentence all (within paragraph) std	0.133291460523
LSA sentence givenness mean	0.559316215663
LSA sentence givenness std	0.0920409453869
LSA sentence span mean	0.61616340573
LSA sentence span std	0.120499279752
Possível classificação do nível escolar	Ensino Fundamental

Figura 5: Resultado do exemplo

Foi adicionado os resultados da *feature* juntamente com as existentes e utilizado o classificador SVM usando 80% de treino e 20% de teste (Tabela 16) foi possível observar uma melhora na classificação. Na Tabela 17 mostra o classificador SVM com validação cruzada usando a *feature* adicional. Note que também há um resultado superior em comparação ao resultado sem a nova *feature*.

Tabela 17: Classificação usando SVM com nova *feature* (*Cross Validation*)

Classe	Precisão	Recall	F-Measure
Ensino Fundamental	0,795	0,850	0,822
Ensino Médio	0,761	0,685	0,721
Média	**0,781**	0,782	0,780

5.4 Corte de *Features*

Durante o desenvolvimento do projeto, procuramos selecionar as *features* mais relevantes para classificação. Para isso, foi utilizado o método de testes estatísticos uni variados, no caso *Analysis of Variance* (ANOVA) usando a biblioteca *sklearn* (*sklearn.feature_selection.f_classif*). Com isso, foi gerado um *ranking* com as *features* mais relevantes. Como descrito na Tabela 18.

Feature	Pontuação
Sobreposição de Radicais de palavras Adjacentes	38.92
Cláusulas por Sentença	35.21
Palavras por Sentenças	34.25
Sobreposição de Radicais de palavras	31.66
Sobreposição de Palavras de conteúdo em sentenças adjacentes	23.31
LSA: média entre sentenças adjacentes	21.29
Complexidade de Yngve	19.47
Sentenças por Parágrafos	19.18
Número de Sentenças	19.18
Sobreposição de Argumentos Adjacentes	19.01
LSA: média entre sentenças todos os pares de sentenças	16.79
Sobreposição de Argumentos	16.71
Distância de dependência	16.09

Tabela 18: *Features* com valor de *F-test* maior que 15

A principal *feature* representativa foi a **Sobreposição de Radicais de palavras Adjacentes** que está no grupo de métricas de Correferência. Sua principal função é calcular uma proporção de sentenças adjacentes que compartilham radicais. Por Exemplo: Dentro do lago, existem peixes, como a traíra e o dourado, além da palometa, um tipo de piranha. Ela é uma espécie carnívora que se alimenta de peixes. No verão, elas ficam mais próximas das margens da barragem, atraídas pela movimentação das pessoas e por restos de comida que alguns turistas deixam na água quando lavam os pratos. Neste exemplo, para sentenças adjacentes, temos que (1) e (2) compartilham o radical "peix". Como há dois pares de sentenças adjacentes e somente um compartilha um radical, então o resultado da métrica é $\frac{1}{2} = 0{,}5$.

Uma outra *feature* que obteve um resultado bastante representativo, com 35.21 de pontuação no ANOVA, foi a **Cláusulas por Sentença**. Essa *feature* calcula o número

médio de cláusulas por sentença. Defini-se que uma cláusula em uma sentença é caracterizada pela presença de um Sintagma Verbal (SV). Por exemplo, "A mulher que eu vi usava um chapéu vermelho. O seu chapéu era muito bonito." A primeira sentença possui 2 cláusulas e a segunda sentença possui uma cláusula. Assim, o valor da métrica é 1,5.

A seleção de *features* é um método capaz de observarmos características específicas nas amostras não visíveis por uma pessoa comum, como no caso das duas *features* descritas, e demonstra ter valores com maior representatividade na separação das amostras comparadas com as outras *features* utilizadas e sendo possível selecioná-las para possível redução da quantidade de *features* do conjunto (Redução de Dimensionalidade).

6 Conclusão

Como foi observado na análise da concordância, essa é uma tarefa de classificação difícil até para humanos. Foi usado os algoritmos *Support Vector Machine* (SVM) e *Perceptron* Multicamadas (MLP), para classificar os documentos entre ensino fundamental e ensino médio. Os resultados foram baixos, com aproximadamente 64% de precisão para o SVM e 60% para o MLP, com a amostra dividida entre 80% treinamento e 20% teste. Usando *cross-validation*, com k=10, o SVM alcançou a precisão de 68%. O valor do k no *cross-validation* foi obtido experimentalmente, a partir de observações com outros valores. Esse resultado foi devido à baixa quantidade de materiais de treinamento. Infelizmente, existem poucos materiais disponíveis para este tipo de tarefa, nenhum corpus foi encontrado com o rótulo.

Além desses testes, foram feitos testes usando uma *feature* adicional, que consiste em calcular a média da profundidade das árvores sintáticas de diferentes sentenças. Com essa nova *feature*, os resultados foram mais promissores. O SVM, sendo testado com a amostra dividida entre treinamento e teste, da mesma forma que foi feita anteriormente, teve precisão de 70%. Usando *cross-validation* a precisão foi de 78%.

Por fim, com este trabalho é possível observar que com a extração de *features* textuais podemos realizar a predição de textos para níveis escolares por meio de classificadores e, considerando o baixo valor de concordância entre os anotadores, notamos que a máquina é mais provável de acerto do que um ser humano.

7. Referências

[1] SCARTON, C.; ALUISIO, S. M. Coh-metrix-port: a readability assessment tool for texts in brazilian portuguese. In: SN. *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings, PROPOR*. [S.l.], 2010. v. 10, n. 1.

[2] CUNHA, A. L. V. d. *Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais*. Tese (Doutorado) — Universidade de São Paulo, 2016.

[3] WIKILIVROS. *Problemas do meio ambiente*. 2007. `https://pt.wikibooks.org/wiki/Problemas_do_meio_ambiente`. Acessado em: 2018-06-17.

[4] FRAZIER, L. Syntactic complexity. *Natural language parsing: Psychological, computational, and theoretical perspectives*, p. 129–189, 1985.

[5] DOWTY, D. R.; KARTTUNEN, L.; ZWICKY, A. M. *Natural language parsing: Psychological, computational, and theoretical perspectives*. [S.l.]: Cambridge University Press, 2005.

[6] SILVA, J. et al. Out-of-the-box robust parsing of portuguese. In: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2010. p. 75–85.

[7] YNGVE, V. H. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, JSTOR, v. 104, n. 5, p. 444–466, 1960.

A Apêndice

A.1 Incidência de Adjetivos

Incidência de adjetivos em um texto. O desempenho da métrica é diretamente relacionado ao desempenho do POS tagger do nlpnet. Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Com 6 adjetivos (polêmico, municipais, estaduais, federais, solares e anuais) e 95 palavras, a incidência de adjetivos é $\frac{\text{número de adjetivos}}{\frac{nmerodepalavras}{1000}} = 63{,}157$.

A.2 Incidência de Advérbios

Incidência de advérbios em um texto. O desempenho da métrica é diretamente relacionado ao desempenho do POS tagger do nlpnet. Exemplo: "Não podemos acrescentar nenhuma despesa a mais no nosso orçamento. Já não temos recursos suficientes para a manutenção das escolas, por exemplo, e também precisamos valorizar o magistério - justifica a diretora do Departamento Pedagógico da SEC, Sonia Balzano."Com 8 advérbios (não, a, mais, já, não, por, exemplo, também) e 38 palavras, a incidência de adjetivos é 210,526 (número de advérbios/(número de palavras/1000)).

A.3 Incidência de Palavras de Conteúdo

Incidência de palavras de conteúdo em um texto. Palavras de conteúdo são substantivos, verbos, adjetivos e advérbios. O desempenho da métrica é diretamente relacionado ao desempenho do POS tagger do nlpnet. Exemplo:"Não podemos acrescentar nenhuma despesa a mais no nosso orçamento. Já não temos recursos suficientes para a manutenção das escolas, por exemplo, e também precisamos valorizar o magistério - justifica a diretora do Departamento Pedagógico da SEC, Sonia Balzano."Com 27 palavras de conteúdo e 38 palavras, a incidência de palavras de conteúdo é 710,526 (número de palavras de conteúdo/(número de palavras/1000)).

A.4 Índice Flesch

O Índice de Legibilidade de Flesch busca uma correlação entre tamanhos médios de palavras e sentenças e a facilidade de leitura. A adaptação do Índice Flesch da língua inglesa para a portuguesa foi realizada por Martins, Teresa B. F., Claudete M. Ghiraldelo, Maria das Graças Volpe Nunes e Osvaldo Novais de Oliveira Junior (1996). Readability formulas applied to textbooks in brazilian portuguese. Notas do ICMC, N. 28, 11p. Fórmula: ILF = 248.835 - [1.015 x (Número de palavras por sentença)] - [84.6 x (Número de sílabas do texto / Número de palavras do texto)] Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Com média de 23 palavras por sentença e 2,31 sílabas por palavra, o índice Flesch para o exemplo é 29,316.

A.5 Incidência de Palavras Funcionais

Incidência de palavras funcionais em um texto. Palavras funcionais são artigos, preposições, pronomes, conjunções e interjeições. O desempenho da métrica é diretamente relacionado ao desempenho do POS tagger do nlpnet. Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Com 26 palavras funcionais e 69 palavras, a incidência de palavras funcionais é 376,81 (número de palavras funcionais/(número de palavras/1000)).

A.6 Sentenças por Parágrafos

Número de sentenças dividido pelo número de parágrafos. Exemplo: "No caso do Jeca Tatu, o verme que o deixou doente foi outro: o Ancylostoma. A larva desse verme vive no solo e penetra diretamente na pele. Só o contrai quem anda descalço na terra contaminada por fezes humanas. Se não se tratar, a pessoa fica fraca, sem ânimo e com a pele amarelada. Daí a doença ser também conhecida como amarelão."O parágrafo do exemplo possui 5 sentenças.

A.7 Sílabas por Palavra de Conteúdo

Número médio de sílabas por palavras de conteúdo (substantivos, verbos, adjetivos e advérbios). O desempenho da métrica é diretamente relacionado ao desempenho do POS tagger do nlpnet. Exemplo: "Acessório utilizado por adolescentes, o boné é um dos itens que compõem a vestimenta idealizada pela proposta."Número de sílabas por palavras de conteúdo do exemplo é 3,5.

A.8 Palavras por Sentenças

Número de palavras dividido pelo número de sentenças. Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns

(PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Neste exemplo o número de palavras é 95 e o número de sentenças é 4. Portanto,o número de palavras por sentenças é 23,75.

A.9 Incidência de Substantivos

Incidência de substantivos em um texto. O desempenho da métrica é diretamente relacionado ao desempenho do POS tagger do nlpnet. Exemplo: "Acessório utilizado por adolescentes, o boné é um dos itens que compõem a vestimenta idealizada pela proposta."Com 6 substantivos (acessório, adolescentes, boné, itens, vestimenta e proposta) e 17 palavras, a incidência de substantivos é 352,94 (número de substantivos/(número de palavras /1000)).

A.10 Número de Parágrafos

Número de parágrafos de um texto. Consideramos como parágrafos somente a quebra de linha (não indentações). Exemplo: "No caso do Jeca Tatu, o verme que o deixou doente foi outro: o Ancylostoma. A larva desse verme vive no solo e penetra diretamente na pele. Só o contrai quem anda descalço na terra contaminada por fezes humanas. Se não se tratar, a pessoa fica fraca, sem ânimo e com a pele amarelada. Daí a doença ser também conhecida como amarelão. Os vermes – também chamados de helmintos – são parasitos, animais que, em geral, dependem da relação com outros seres para viver. Eles podem se hospedar no organismo de diversos animais, como bois, aves e peixes. Por isso, podemos também contraí-los comendo carnes cruas ou mal cozidas."O exemplo possui 2 parágrafos.

A.11 Número de Sentenças

Número de sentenças de um texto. Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."O exemplo possui 4 sentenças.

A.12 Número de Palavras

Número de palavras do texto. Exemplo: "Acessório utilizado por adolescentes, o boné é um dos itens que compõem a vestimenta idealizada pela proposta."O exemplo possui 17 palavras.

A.13 Incidência de Pronomes

Incidência de pronomes em um texto. Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Com 2 pronomes (quem e ele) e 69 palavras, a incidência de pronomes é 28,98 (número de pronomes/(número de palavras/1000)).

A.14 Incidência de Verbos

Incidência de verbos em um texto. O desempenho da métrica é diretamente relacionado ao desempenho do POS tagger do nlpnet. Exemplo: "Acessório utilizado por adolescentes, o boné é um dos itens que compõem a vestimenta idealizada pela proposta."Com 4 verbos (utilizado, é, compõem e idealizada)e 17 palavras, a incidência de verbos é 235,29 (número de verbos/(número de palavras /1000)).

A.15 Incidência do operador lógico E

Incidência do operador lógico E em um texto. Exemplo: "Não podemos acrescentar nenhuma despesa a mais no nosso orçamento. Já não temos recursos suficientes para a manutenção das escolas, por exemplo, e também precisamos valorizar o magistério - justifica a diretora do Departamento Pedagógico da SEC, Sonia Balzano."Como há 1 operadores lógicos E e 38 palavras a incidência do operadores lógico E é 26,315 (frequência do operador lógico E / (número de palavras/1000)).

A.16 Incidência do operador lógico SE

Incidência do operador lógico SE em um texto (desconsidera quando o SE é um pronome). Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Como há 1 operadores lógicos SE e 95 palavras a incidência do operadores lógico SE é 10,526 (frequência do operador lógico SE / (número de palavras/1000)).

A.17 Incidência do operador lógico OU

Incidência do operador lógico OU em um texto. Exemplo: "Os vermes – também chamados de helmintos – são parasitos, animais que, em geral, dependem da relação com outros seres para viver. Eles podem se hospedar no organismo de diversos animais, como bois, aves e peixes. Por isso, podemos também contraí-los comendo carnes cruas ou mal cozidas."Como há 1 operadores lógicos OU e 45 palavras a incidência do operadores lógico OU é 22,222 (frequência do operador lógico OU / (número de palavras/1000)).

A.18 Incidência de negação

Incidência de Negações. Consideramos como negações: não, nem, nenhum, nenhuma, nada, nunca e jamais. Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."No exemplo aparecem 3 negações. Como o mesmo possui 38 palavras a incidência de negações é 78,947 (número de negações/(número de palavras/1000)).

A.19 Incidência de Operadores Lógicos

Incidência de operadores lógicos em um texto. Consideramos como operadores lógicos: e, ou, se, negações e um número de condições. Exemplo: "Não podemos acrescentar nenhuma despesa a mais no nosso orçamento. Já não temos recursos suficientes para a manutenção das escolas, por exemplo, e também precisamos valorizar o magistério - justifica a diretora do Departamento Pedagógico da SEC, Sonia Balzano."Como há 4 operadores lógicos e 38 palavras a incidência de operadores lógicos é 105,26 (número de operadores lógicos/(número de palavras/1000)).

A.20 Frequência das palavras de conteúdo

Média de todas as frequências das palavras de conteúdo (substantivos, verbos, advérbios e adjetivos) encontradas no texto. O valor da frequência das palavras é retirado da lista de frequências do córpus Banco de Português (BP), compilado por Tony Sardinha da PUC-SP. O desempenho da métrica é diretamente relacionado a lista de frequências compilada do córpus BP. Exemplo: "Acessório utilizado por adolescentes, o boné é um dos itens que compõem a vestimenta idealizada pela proposta."O texto possui as palavras de conteúdo: Acessório, utilizado, adolescentes, boné, é, itens, compõem, vestimenta, idealizada, proposta; com frequências 1616, 78716, 53937, 1615, 5325656, 32350, 17961, 773, 1908, 135451. O valor da métrica é 564998.3

A.21 Frequência da palavra de conteúdo mais rara

Primeiramente identificamos a menor frequência dentre todas as palavras de conteúdo (substantivos, verbos, advérbios e adjetivos) em cada sentença. Depois, calculamos uma média de todas as frequências mínimas. A palavra com a menor frequência é a mais rara da sentença. O desempenho da métrica é diretamente relacionado a lista de frequências compilada do córpus BP. Exemplo: "Acessório utilizado por adolescentes, o boné é um dos itens que compõem a vestimenta idealizada pela proposta."O texto possui as palavras de conteúdo: Acessório, utilizado, adolescentes, boné, é, itens, compõem, vestimenta, idealizada, proposta. A menor frequência é de vestimenta, 773.

A.22 Hiperônimos de Verbos

Para cada verbo soma-se o número de hiperônimos e divide o total pelo número de verbos. Hiperonímia é uma relação, definida na Wordnet.Br (Dias-da-Silva et. al., 2002; Dias-da-Silva, 2003; Dias-da-Silva, 2005; Dias-da-Silva et. al., 2008 e Scarton e Aluísio, 2009), de "super tipo de". O verbo sonhar, por exemplo, possui 3 hiperônimos: imaginar, conceber e ver na mente. O desempenho da métrica é diretamente relacionado ao desempenho da base Wordnet.Br. Exemplo: "Ele sonha muito quando está acordado."O verbo sonhar possui 3 hiperônimos e o verbo acordar nenhum. Assim, temos o valor de 1,5.

A.23 Índice de Brunet

W=N $\times$ ($V \times 0.165$) N é o número de palavras lexicais, e V é o número total de tokens usados. Os valores de W típicos variam entre 10 e 20, sendo que uma fala mais rica produz valores menores (THOMAS et al., 2005).

Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."

Com 95 tokens e 78 tipos, a métrica vale 9,199.

A.24 Estatística de Horoné

A Estatística de Honoré R, calculada como: $R = 100 * logN/(1 - (V_1/V))$ em que N é o número total de tokens, V_1 é o número de palavras do vocabulário que aparecem uma única vez, e V é o número de palavras lexicais.

Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."

Com 95 tokens, 69 tokens com apenas uma ocorrência e 78 tipos, o valor da métrica é 1714,027.

A.25 Cláusulas por Sentença

Calcula o número médio de cláusulas por sentença. Definiu-se que uma cláusula em uma sentença é caracterizada pela presença de um sintagma verbal.

Exemplo: "A mulher que eu vi usava um chapéu vermelho. O seu chapéu era muito bonito."

A primeira sentença possui 2 cláusulas e a segunda sentença possui uma cláusula. Assim, o valor da métrica é 1,5.

A.26 Incidência de Pronomes Pessoais

Incidência de pronomes pessoais em um texto. Consideramos como pronomes pessoais: eu, tu, ele/ela, nós, nós, eles/elas, você e vocês. Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Este exemplo possui 1 pronome pessoal. Como este texto possui 95 palavras, a incidência de pronomes pessoais é 10,526 (número de pronomes pessoais/(número de palavras/1000)).

A.27 Pronomes por Sintagmas

Média do número de pronomes que aparecem em um texto pelo número de sintagmas nominais. O desempenho dessa métrica é diretamente relacionado ao desempenho das árvores sintáticas de constituintes geradas pelo LX-Parser [1]. [1] Silva, João, António Branco, Sérgio Castro e Ruben Reis. Out-of-the-Box Robust Parsing of Portuguese. In Proceedings of the 9th International Conference on the Computational Processing of Portuguese (PROPOR'10), pp. 75–85. Exemplo: "Dentro do lago, existem peixes, como a traíra e o dourado, além da palometa, um tipo de piranha. Ela é uma espécie carnívora que se alimenta de peixes."Não há pronomes na primeira sentença e há 9 sintagmas nominais. Há 1 pronome na segunda sentença e 5 sintagmas nominais. Com 1 pronome em 2 sentenças, o valor da métrica é 0,1.

A.28 Relação Tipo por Token (Type)

Número de palavras únicas dividido pelo número de tokens dessas palavras. Cada palavra única é um tipo. Cada instância desta palavra é um token. Por exemplo, se a palavra cachorro aparece 7 vezes em um texto, seu tipo (type) é 1 e seu token é 7. Calculamos esta métrica somente para palavras de conteúdo (substantivos, verbos, advérbios e adjetivos). Observação: Não usamos lematização de palavras, ou seja, a palavra cachorro é considerada diferente de cachorros. Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Com 95 tokens e 78 tipos, a relação tipo por token é 0,821.

A.29 Incidência de Sintagmas (Constituintes)

Incidência de sintagmas nominais por 1000 palavras. O desempenho da métrica é diretamente relacionada as árvores sintáticas de constituintes geradas pelo LX-Parser. Exemplo: "Acessório utilizado por adolescentes, o boné é um dos itens que compõem a

vestimenta idealizada pela proposta."Como o texto possui 5 sintagmas nominais e 17 palavras a incidência de sintagmas é 294,11 (número de sintagmas/(número de palavras/1000)).

A.30 Palavras Antes de Verbos Principais (Constituintes)

Média de palavras antes de verbos principais na cláusula principal da sentença. Segundo a documentação do Coh-Metrix é um bom índice para avaliar a carga da memória de trabalho. O desempenho da métrica é diretamente relacionada às árvores sintáticas de dependência geradas pelo MaltParser e ao POS tagger do nlpnet. Exemplo: "Acessório utilizado por adolescentes, o boné é um dos itens que compõem a vestimenta idealizada pela proposta."Como este texto possui uma sentença o valor desta métrica corresponde ao valor de palavras antes do verbo desta única sentença que, neste caso, é 1 (a palavra acessório é a única que antecede o verbo).

A.31 Incidência de Conectivos

Incidência de todos os conectivos que aparecem em um texto. Para esta métrica (e as demais que contam conectivos) compilamos listas de conectivos classificados em duas dimensões. A primeira dimensão divide os conectivos em positivos e negativos (conectivos positivos estendem eventos, enquanto que conectivos negativos param eventos). A segunda dimensão divide os conectivos de acordo com o tipo de coesão: aditivos, temporais, lógicos e causais. Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Como há 6 conectivos e 95 palavras, a incidência de conectivos é 63,157 (número de conectivos/(número de palavras/1000)).

A.32 Incidência de conectivos classificados como aditivos negativos

Incidência de todos os conectivos negativos que aparecem em um texto. Exemplo: "Entretanto, foram encontrados vários problemas clássicos."Como há 1 conectivos negativos (entretanto) e 6 palavras, a incidência de conectivos negativos é 166,666 (número de conectivos positivos/(número de palavras/1000)).

A.33 Incidência de conectivos classificados como aditivos positivos

Incidência de todos os conectivos positivos que aparecem em um texto. Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo

excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Como há 5 conectivos positivos e 95 palavras, a incidência de conectivos positivos é 52,631 (número de conectivos positivos/ (número de palavras/1000)).

A.34 Incidência de conectivos classificados como causais negativos

Incidência de todos os conectivos causais negativos que aparecem em um texto. Exemplo: "Embora tenha colado na prova, o menino não obteve uma boa nota."Como há 1 conectivos causal negativo (Embora) e 12 palavras, a incidência de conectivos causais negativos é 83,333 (número de conectivos causais negativos/(número de palavras/1000)).

A.35 Incidência de conectivos classificados como causais positivos

Incidência de todos os conectivos causais positivos que aparecem em um texto. Exemplo: "O menino queria ir bem na prova. Para isso, ele resolveu colar."Como há 1 conectivos causal positivo (Para isso) e 12 palavras, a incidência de conectivos causais positivos é 83,333 (número de conectivos causais positivos/(número de palavras/1000)).

A.36 Conectivos Lógicos Negativos

Incidência de todos os conectivos lógicos negativos que aparecem em um texto. Exemplo: "O menino colou na prova, embora soubesse que poderia ser pego."Como há 1 conectivos lógico negativo (Desde que) e 11 palavras, a incidência de conectivos lógicos negativos é 90,909 (número de conectivos lógicos negativos/(número de palavras/1000)).

A.37 Conectivos Lógicos Positivos

Incidência de todos os conectivos lógicos positivos que aparecem em um texto. Exemplo: "Desde que o menino começou a colar nas provas, ele não estuda mais."Como há 1 conectivos lógico positivo (Desde que) e 13 palavras, a incidência de conectivos lógicos positivos é 76,923 (número de conectivos lógicos positivos/(número de palavras/1000)).

A.38 Incidência de conectivos classificados como temporais negativos

Incidência de todos os conectivos temporais negativos que aparecem em um texto. Exemplo: "O menino colou na prova até que a professora descobriu sua artimanha."Como há 1 conectivos temporais negativo (até) e 12 palavras, a incidência de conectivos temporais negativos é 83,333 (número de conectivos temporais negativos/(número de palavras/1000)).

A.39 Incidência de conectivos classificados como temporais positivos

Incidência de todos os conectivos temporais positivos que aparecem em um texto. Exemplo: "Enquanto isso, mais de 100 pessoas tentam resolver o problema, o que finalmente resultou em bons resultados."Como há 2 conectivos temporais positivos (enquanto e finalmente) e

6 palavras, a incidência de conectivos temporais positivos é 117,647 (número de conectivos temporais positivos/(número de palavras/1000)).

A.40 Ambiguidade de Adjetivos

Para cada adjetivo do texto soma-se o número de sentidos apresentados no TEP (Maziero et. al., 2008) e divide o total pelo número de adjetivos. O desempenho da métrica é diretamente relacionado ao desempenho do dicionário do TEP. Exemplo: "O acessório polêmico entrou no projeto, de autoria do senador Cícero Lucena (PSDB-PB), graças a uma emenda aprovada na Comissão de Educação do Senado em outubro. Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."Dos adjetivos rotulados no texto exemplo (polêmico, municipal, estadual, federal, solar, anual), consta apenas anual no TEP. Assim, o valor da métrica é 0,166.

A.41 Ambiguidade de Advérbios

Para cada advérbio do texto soma-se o número de sentidos apresentados no TEP (Maziero et. al., 2008) e divide o total pelo número de advérbios. O desempenho da métrica é diretamente relacionado ao desempenho do dicionário do TEP. Exemplo: "Não podemos acrescentar nenhuma despesa a mais no nosso orçamento. Já não temos recursos suficientes para a manutenção das escolas, por exemplo, e também precisamos valorizar o magistério - justifica a diretora do Departamento Pedagógico da SEC, Sonia Balzano."Os advérbios rotulados no texto exemplo são: não, mais, já, não com sentidos: 1, 5, 4, 1. Assim, o valor da métrica é 2,2.

A.42 Ambiguidade de Substantivos

Para cada substantivo do texto soma-se o número de sentidos apresentados no TEP (Maziero et. al., 2008) e divide o total pelo número de substantivos. O desempenho da métrica é diretamente relacionado ao desempenho do dicionário do TEP. Exemplo: "O menino colou na prova, embora soubesse que poderia ser pego."O exemplo apresenta 2 substantivos (menino e prova) com frequências 1 e 9 no TEP. O resultado da métrica é 5,0.

A.43 Ambiguidade de Verbos

Para cada verbo do texto soma-se o número de sentidos apresentados no TEP (Maziero et. al., 2008) e divide o total pelo número de verbos. O desempenho da métrica é diretamente relacionado ao desempenho do dicionário do TEP. Exemplo: "O menino colou na prova, embora soubesse que poderia ser pego."O exemplo apresenta 4 verbos (colou, soubesse, poderia e ser) com frequências 4, 7, 2 e 12 no TEP. O resultado da métrica é 6,25.

A.44 Distância de Dependência

Média da distância de dependência das sentenças de um texto. Para cada sentença do texto, a distância de dependência é calculada como a soma da distância entre as palavras associadas

a cada relação de dependência. As árvores de dependência são extraídas pelo MaltParser Exemplo: Maria foi ao mercado. No mercado, comprou ovos e pão. Para a primeira sentença do exemplo, as relações de dependência e suas respectivas distâncias associadas são:

1. nsubj(foi, Maria), 1
2. adpmod(foi, ao), 1
3. adpobj(ao, mercado), 1

Para esta sentença, a distância de dependência é 3. Para a segunda sentença, as relações de dependência são:

1. adpmod(comprou, No), 2
2. adpobj(No, mercado), 1
3. dobj(comprou, ovos), 1
4. cc(ovos, e), 1
5. conj(ovos, pão), 2

Para esta sentença, a distância de dependência é 7. Portanto, para o exemplo todo, o valor da métrica será (3 + 7) / 2 = 5,0.

A.45 Complexidade de Frazier

Média da complexidade sintática de Frazier, descrita em [4] [5]. Para um dado texto, calcula-se a complexidade sintática de cada sentença, e então calcula-se a média desses valores. No cálculo da complexidade, utiliza-se o máximo da soma de trigramas dos escores das palavras. As árvores sintáticas utilizadas são geradas pelo LX-Parser [6]. Exemplo: Maria foi ao mercado. No mercado, comprou ovos e pão. No exemplo, a primeira sentença possui complexidade de Frazier 6,0, e a segunda, 5,0, para um valor total de (6,0 + 5,0) / 2 = 5,5.

A.46 Complexidade de Yngve

Média da complexidade sintática de Yngve, descrita em [7]. Para um dado texto, calcula-se a complexidade sintática de cada sentença, e então calcula-se a média desses valores. No cálculo da complexidade, utiliza-se a média dos escores das palavras. As árvores sintáticas utilizadas são geradas pelo LX-Parser [6]. Exemplo: Maria foi ao mercado. No mercado, comprou ovos e pão. No exemplo, a primeira sentença possui complexidade de Yngve 1,4, e a segunda, 2,0, para um valor total de (1,4 + 2,0) / 2 = 1,7.

A.47 Sobreposição de Argumentos Adjacentes (Correferência)

Proporção de sentenças adjacentes que compartilham um ou mais argumentos (substantivos, pronomes ou sintagmas nominais). O desempenho da métrica é diretamente relacionada ao do POS tagger do nlpnet. Exemplo: Dentro do lago, existem peixes, como a traíra e o dourado, além da palometa, um tipo de piranha. Ela é uma espécie carnívora que se

alimenta de peixes. No verão, elas ficam mais próximas das margens da barragem, atraídas pela movimentação das pessoas e por restos de comida que alguns turistas deixam na água quando lavam os pratos. Para este exemplo temos que as sentenças (1) e (2) compartilham um substantivo (peixes) e, portanto, este par incrementa 1 no valor de correferência. Como também há dois pares de sentenças adjacentes ((1) com (2) e (2) com (3)), o valor final da métrica é 1/2 = 0,5.

A.48 Sobreposição de Argumentos (Correferência)

Proporção de todos os pares de sentenças que compartilham um ou mais argumentos (substantivos, pronomes ou sintagmas nominais). O desempenho da métrica é diretamente relacionada ao do POS tagger do nlpnet. Exemplo: Dentro do lago, existem peixes, como a traíra e o dourado, além da palometa, um tipo de piranha. Ela é uma espécie carnívora que se alimenta de peixes. No verão, elas ficam mais próximas das margens da barragem, atraídas pela movimentação das pessoas e por restos de comida que alguns turistas deixam na água quando lavam os pratos. Para este exemplo temos os pares de sentenças (1) com (2), (1) com (3) e (2) com (3). Como somente o par (1) com (2) compartilham um substantivo (peixes) o valor final da métrica é 1/3 = 0,333.

A.49 Sobreposição de Radicais de palavras Adjacentes (Correferência)

Proporção de sentenças adjacentes que compartilham radicais. Exemplo: Dentro do lago, existem peixes, como a traíra e o dourado, além da palometa, um tipo de piranha. Ela é uma espécie carnívora que se alimenta de peixes. No verão, elas ficam mais próximas das margens da barragem, atraídas pela movimentação das pessoas e por restos de comida que alguns turistas deixam na água quando lavam os pratos. Neste exemplo, para sentenças adjacentes, temos que (1) e (2) compartilham o radical "peix". Como há dois pares de sentenças adjacentes e somente um compartilha um radical, então o resultado da métrica é 1/2 = 0,5.

A.50 Sobreposição de Radicais de palavras (Correferência)

Proporção de todos os pares de sentenças que compartilham radicais. Exemplo: Dentro do lago, existem peixes, como a traíra e o dourado, além da palometa, um tipo de piranha. Ela é uma espécie carnívora que se alimenta de peixes. No verão, as piranhas ficam mais próximas das margens da barragem, atraídas pela movimentação das pessoas e por restos de comida que alguns turistas deixam na água quando lavam os pratos. Neste exemplo, além de (1) e (2), (1) e (3) também compartilham um radical (piranh). Então, como são três os possíveis pares, o valor final da métrica é 2/3 = 0,667.

A.51 Sobreposição de Palavras de conteúdo em sentenças adjacentes (Correferência)

Proporção de sentenças adjacentes que compartilham palavras de conteúdo (substantivos, verbos, adjetivos e advérbios). O desempenho da métrica é diretamente relacionada ao do POS tagger do nlpnet. Exemplo: Dentro do lago, existem peixes, como a traíra e o dourado, além da palometa, um tipo de piranha. Ela é uma espécie carnívora que se alimenta de

peixes. No verão, as piranhas ficam mais próximas das margens da barragem, atraídas pela movimentação das pessoas e por restos de comida que alguns turistas deixam na água quando lavam os pratos. Neste exemplo, além de (1) e (2), (1) e (3) também compartilham um radical (piranh). Então, como são três os possíveis pares, o valor final da métrica é 2/3 = 0,667.

A.52 Referência Anafórica Adjacente

Proporção de referências anafóricas entre sentenças adjacentes. Exemplo: "Dentro do lago, existem peixes, como a traíra e o dourado, além da palometa, um tipo de piranha. Ela é uma espécie carnívora que se alimenta de peixes."Com referências anafóricas de "Ela"para "traíra"e "piranha", temos 2 candidatos para 1 referência. Logo, a métrica retorna 2. obs: "palometa"não é reconhecida no dicionário UNITEX.

A.53 Referência Anafórica

Proporção de referências anafóricas que se referem a um constituinte presente em até cinco sentenças anteriores. Exemplo: "Dentro do lago, existem peixes, como a traíra e o dourado, além da palometa, um tipo de piranha. Ela é uma espécie carnívora que se alimenta de peixes."Com referências anafóricas Ela para traíra e piranha, temos 2 candidatos para 1 referência. Logo, a métrica retorna 2. Obs: palometa não é reconhecida no dicionário UNITEX.

A.54 LSA: média entre sentenças adjacentes

Média de similaridade entre pares de sentenças adjacentes no texto. O espaço LSA utilizado na versão atual do sistema foi gerado a partir do mesmo corpus empregado na geração do modelo de língua utilizado pela métrica de entropia cruzada: um corpus de 120.813.620 tokens, que consiste na união dos corpora Wikipedia, PLN-BR, LácioWeb, e Revista Pesquisa FAPESP. Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."O exemplo possui 3 sentenças, e, portanto, 2 pares de sentenças adjacentes. A similaridade LSA entre a primeira e a segunda sentenças, segundo o modelo utilizado na versão atual do Coh-Metrix-Dementia, é 0,084, e a similaridade entre a segunda e a terceira sentenças é 0,063. Nesse caso, a média entre esses valores é de 0,0735.

A.55 LSA: desvio padrão entre sentenças adjacentes

Desvio padrão de similaridade entre pares de sentenças adjacentes no texto. Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."O exemplo possui 3 sentenças, e, portanto, 2 pares de sentenças adjacentes. A similaridade LSA entre a primeira

e a segunda sentenças é 0,084, e a similaridade entre a segunda e a terceira sentenças é 0,063. O desvio padrão entre esses valores é de 0,0105.

A.56 LSA: média entre sentenças todos os pares de sentenças

Média de similaridade entre todos os pares de sentenças no texto. Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."O exemplo possui 3 sentenças, e, portanto, 3 pares de sentenças. A similaridade LSA entre a primeira e a segunda sentenças é 0,084, a similaridade entre a segunda e a terceira sentenças é 0,063, e a similaridade entre a primeira e a terceira é 0,362. A média entre esses valores é 0,17.

A.57 LSA: desvio padrão entre sentenças todos os pares de sentenças

Desvio padrão de similaridade entre todos os pares de sentenças no texto. Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."O exemplo possui 3 sentenças, e, portanto, 3 pares de sentenças. A similaridade LSA entre a primeira e a segunda sentenças é 0,084, a similaridade entre a segunda e a terceira sentenças é 0,063, e a similaridade entre a primeira e a terceira é 0,362. O desvio padrão entre esses valores é 0,14.

A.58 LSA: média entre parágrafos adjacentes

Média de similaridade entre pares de parágrafos adjacentes no texto. Esta métrica é calculada do mesmo modo que a média entre sentenças adjacentes, mas utilizando-se parágrafos, ao invés de sentenças, como unidades.

A.59 LSA: desvio padrão entre parágrafos adjacentes

Desvio padrão de similaridade entre pares de parágrafos adjacentes no texto. Esta métrica é calculada do mesmo modo que o desvio padrão entre sentenças adjacentes, mas utilizando-se parágrafos, ao invés de sentenças, como unidades.

A.60 LSA: média de givenness das sentenças

Média do givenness da cada sentença do texto, a partir da segunda. Se o texto possui apenas uma sentença, define-se a métrica como 0,0. Define-se o givenness de uma sentença como a similaridade LSA entre a sentença e todo o texto que a precede. Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a

medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."O exemplo possui 3 sentenças. A similaridade LSA entre a primeira e a segunda sentenças é 0,084, e a similaridade entre a terceira e o conjunto formado pela primeira e a segunda é 0,286. A média entre esses valores é 0,185.

A.61 LSA: desvio padrão de givenness das sentenças

Desvio padrão do givenness da cada sentença do texto, a partir da segunda. Se o texto possui apenas uma sentença, define-se a métrica como 0,0. Define-se o givenness de uma sentença como a similaridade LSA entre a sentença e todo o texto que a precede. Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."O exemplo possui 3 sentenças. A similaridade LSA entre a primeira e a segunda sentenças é 0,084, e a similaridade entre a terceira e o conjunto formado pela primeira e a segunda é 0,286. O desvio padrão entre esses valores é 0,101.

A.62 LSA: média do span das sentenças

Média do span da cada sentença do texto, a partir da segunda. Se o texto possui apenas uma sentença, define-se a métrica como 0,0. O span de uma sentença, assim como o givenness, é uma forma de medir a proximidade entre uma sentença e o contexto que a precede. A diferença, em termos simples, consiste no fato de que o span procura capturar a similaridade não apenas com o conteúdo explícito apresentado anteriormente no texto, mas também com tudo o que se pode inferir com base nesse conteúdo. Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."O exemplo possui 3 sentenças. O span LSA entre a primeira e a segunda sentenças, segundo o modelo utilizado na versão atual do Coh-Metrix-Dementia, é 0,084, e o span entre a terceira e o conjunto formado pela primeira e a segunda é 0,223. A média desses valores é 0,1535.

A.63 LSA: desvio padrão do span das sentenças

Desvio padrão do span da cada sentença do texto, a partir da segunda. Se o texto possui apenas uma sentença, define-se a métrica como 0,0. Exemplo: "Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a idéia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta."O exemplo possui 3 sentenças.

O span LSA entre a primeira e a segunda sentenças é 0,084, e o span entre a terceira e o conjunto formado pela primeira e a segunda é 0,223. O desvio padrão desses valores é 0,070.

A.64 Densidade de Conteúdo

A densidade de conteúdo de um texto é calculada como o número de palavras de classe aberta (também denominadas palavras de conteúdo) dividido pelo número de palavras de classe fechada (ou palavras funcionais). Exemplo: Maria foi ao mercado. No mercado, comprou ovos e pão. No exemplo, há 7 palavras de conteúdo (Maria, foi, mercado, mercado, comprou, ovos, pão), e 3 palavras funcionais (ao, no, e), o que resulta num valor de 7/3 = 2,33.