UNIVERSIDADE DE ITAÚNA
PRÓ-REITORIA DE ENSINO
COORDENAÇÃO DE CIÊNCIA DA COMPUTACAO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Eugênio Cunha

APRENDIZADO DE MÁQUINA APLICADO À VALORAÇÃO DE REDAÇÕES

ITAÚNA
2017

EUGÊNIO CUNHA

APRENDIZADO DE MÁQUINA APLICADO À VALORAÇÃO DE REDAÇÕES

Projeto submetido à Coordenadoria do Curso de Bacharelado em Ciência da Computação da Universidade de Itaúna - Campus Verde, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Área de pesquisa: Aprendizagem de Máquina

Orientador: Prof. Dr. Marco Túlio Alves N Rodrigues

Itaúna
2017

UNIVERSIDADE DE ITAÚNA
COORDENAÇÃO DE CIÊNCIA DA COMPUTAÇÃO

EUGÊNIO CUNHA

Este projeto foi julgada adequada para a obtenção do Grau de Bacharel em Ciência da Computação, sendo aprovada pela coordenação de ciência da computação do curso de Bacharelado em Ciência da Computação do Campus verde da Universidade de Itaúna e pela banca examinadora:

_____

Orientador: Prof. Dr. Marco Túlio Alves N Rodrigues
Universidade de Itaúna- UIT

_____

Avaliador: Prof. Me. Felipe Domingos da Cunha
Universidade de Itaúna- UIT

Itaúna, 19 de Junho de 2017

Dedico este trabalho a Davi, meu filho, sempre preocupado em proporcionar um minuto de pausa para brincadeiras durante minhas horas de estudo, “meus melhores minutos”!

“Os computadores são incrivelmente rápidos, precisos e burros; os homens são incrivelmente lentos, imprecisos e brilhantes; juntos, seus poderes ultrapassam os limites da imaginação”.

Albert Einstein

Resumo

Este trabalho baseou-se na avaliação das competências exigidas em um texto de redação do tipo dissertativo-argumentativo com temas diversificados de ordem social, científica, cultural ou política. Fundamentou-se no estudo das técnicas de aprendizado de máquina supervisionado que provê uma gama diversificada de algoritmos poderosos para classificações de textos.

O objetivo deste trabalho é classificar as competências exigidas em um texto de redação a partir do treinamento de um modelo de aprendizado de máquina com base em um *corpus* de redações avaliadas seguindo as competências exigidas em uma redação do tipo dissertativa-argumentativa.

A compilação de um *corpus* de redações para treinamento e teste de um modelo de aprendizado de máquina exigiu a prática de extração de informações que compreende técnicas e algoritmos que realisam duas tarefas importantes: a identificação de informações desejadas a partir de documentos estruturados e não-estruturados, e o armazenamento dessas informações em um formato apropriado para uso futuro. Afim de se avaliar a eficácia dos classificadores, vários experimentos foram executados usando-se um *corpus* extraido do banco de redações da UOL (UOL, 2017), um serviço que estimula o estudante a treinar produção de textos, em especial do gênero dissertativo-argumentativo.

O resultado de classificação das competências exigidas em um texto de redação obtidas experimentalmente mostraram que o resultado geral do sistema proposto é comparável à avaliação manual por avaliadores capacitados.

Palavras-chaves: Aprendizado de máquina, Banco de Redações, Classificação, Redação

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

TAG	Rótulo
BOW	Representação da frequência acumulada de palavras em documentos diferentes a partir de um dicionário pré-criado

Lista de Abreviacoes

CEBRASPE	Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos
CSF	Ciência sem fronteiras
ENEM	Exame Nacional do Ensino Médio
HTML	*HyperText Markup Language*
IA	Inteligência artificial
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais
SISU	Sistema de Seleção Unificada
UNB	Universidade de Brasília
JSON	*JavaScript Object Notation*

CAPÍTULO 1

INTRODUÇÃO

O desenvolvimento de uma redação e uma atividade prática presente na cultura civilizada desde a invenção da escrita. Já faz pelo menos uma década que um bom desempenho na redação do Exame Nacional de Ensino Médio - ENEM virou sinônimo de chances maiores para ser aprovado no processo seletivo de acesso a inúmeras universidades públicas (SISU, 2017) e a importantes programas de governo como Ciência sem fronteiras (CSF, 2017).

Em todo processo seletivo é comum o uso de marcações em gabaritos afim de automatizar o processo de correção, uma alternativa rápida e segura, até mesmo aplicações de provas eletrônicas são cada vez mais comum. Um exemplo seria o processo seletivo para avaliador das redações do ENEM que durante a "FASE II" respondem uma prova eletrônica eliminatória (CEBRASPE, 2016a). É notável que todo o processo evoluiu com objetivo de agilidade, confiança e segurança do resultado, entretanto a avaliação das competências de uma redação ainda depende exclusivamente da supervisão de duas ou mais pessoas envolvidas (INEP, 2016).

A redação é aplicada no ENEM desde a primeira edição 1998, hoje o maior exame do Brasil, que na edição de 2016 teve 8.627.195 escritos confirmados, e a participação direta de 11.360 profissionais externos na correção de 5.825.134 redações, entre eles, 378 supervisores e 10.982 avaliadores de acordo com a CEBRASPE (CEBRASPE, 2016b).

Segundo o edital do ENEM 2016 (INEP, 2016) cada redação foi avaliada por, pelo menos, dois avaliadores, de forma independente, contabilizando um número mínimo de 11.650.268 avaliações manuais, das competências exigidas em um texto de redação pelo ENEM.

Machine Learning ou aprendizado de máquina é uma área de IA cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Os

algoritmos de aprendizado de máquina procuram padrões dentro de um conjunto de dados (MITCHELL, 1997). Esses algoritmos existem há bastante tempo, uma ciência que não é nova, mas que está ganhando um novo impulso enquanto o processamento computacional cresce e fica mais barato.

A hipótese desta monografia é que um ou mais modelos de aprendizado de máquina na valoração das competências de uma redação pode ser tão eficiênte quanto o processo de avaliação manual.

1.1 Definição do Problema de Pesquisa

Dado um corpus de redações avaliar as competências exigidas em um texto de redação do tipo dissertativo-argumentativo substituindo a etapa de avaliação manual.

1.2 Motivação

Com crescente volume e variedade de dados disponíveis, o processamento computacional que está mais barato e mais poderoso, e o armazenamento de dados de forma acessível, o aprendizado de máquina está no centro de muitos avanços tecnológicos atingindo áreas antes exclusivas de seres humanos. Os carros autônomos do Google são o exemplo de uma atividade antes exclusiva de um humano e hoje exercida e aperfeiçoada por algoritmos de aprendizado de máquina (WAYMO, 2017).

Aplicações de aprendizado de máquina estão presentes na nossa vida cotidiana como, resultados de pesquisa web, análise de sentimento baseado em texto e na detecção de fraudes em operações com cartões de crédito (BATISTA *et al.*, 1999).

As competências exigidas em uma redação podem ser avaliadas por aprendizado de máquina, diferente de um ser humano um algoritmo de aprendizado de máquina está livre de ansiedade, fadiga, *stress* entre outro fatores emocionais que afetão uma avaliação imparcial a opinião do autor.

1.3 Objetivos Gerais e Específicos

Este trabalho tem como objetivo geral aplicar aprendizado de máquina na avaliação das competências exigidas em um texto de redação do tipo dissertativo-argumentativo.

1.3.1 Objetivos Especificos

O método de construção do conhecimento deste trabalho terá como fundamentos processos de pesquisas relacionadas às áreas descritas. O mesmo será dividido em etapas dentro do escopo geral de forma detalhada e refinada para alcançar o objetivo geral acima, são particularizadas como os seguintes objetivos específicos:

- Percorrer o banco de redações UOL (UOL, 2017) em páginas HTML, filtrar e coletar redações avaliadas;
- Normalizar os textos coletados, separar o tema, título, texto e competências avaliadas em uma estrura no formato JSON e armazenar;
- Montar um fluxo de trabalho utilizando a ferramenta para mineração de dados *Orange* (LJUBLJANA, 2017) com modelos classificadores de multiplas classes;
- Ajustar e treinar os modelos classificadores com o corpus de redações;
- Realizar testes de acurácia, *overfitting* e *noise* sobre os classificadores;
- Representar e comparar graficamente os resultados obtidos.

1.4 Contribuições

O presente estudo contribuirá na área do aprendizado de máquina e diretamente no processo de avaliação de um texto em prosa do tipo dissertativo argumentativo.

1.5 Organização do trabalho

Capitulo 2: Trabalhos Relacionados cita alguns dos trabalhos lidos para embasamento teórico que serviram de base para solucionar o problema proposto.

Capitulo 4: Método proposto apresenta as etapas passo a passo para desenvolver e resolver o problema proposto deste trabalho.

Capitulo 5: Desenvolvimento descreve cada procedimento metodológico que será utilizado para a realização da pesquisa.

Capitulo 6: Resultados Experimentais apresenta os resultados obtidos do trabalho desta pesquisa.

CAPÍTULO 2

TRABALHOS RELACIONADOS

2.1 Competências de uma redação

2.2 Ferramentas para mineração de dados

2.3 Modelo adaboost

CAPÍTULO 3

FUNDAMENTAÇÃO TEÓRICA

CAPÍTULO 4

MÉTODO PROPOSTO

Para concluir com êxito o desenvolvimento deste trabalho e consequentemente os objetivos propostos, o método utilizado para solução do problema é composto das seguintes etapas sequenciais:

Como já foi dito o banco de redações UOL foi desenvolvido e armazenado em páginas HTML, o que permite o uso de um *Web Crawler*, um algoritmo que explora a estrutura de grafo da *Web* para navegar de uma página para outra. A Figura 1 ilustra a etapa que o *Web Crawler* recupera as páginas, filtra as redações avaliadas e coleta cada uma para um repositório local.

Figura 1: Um *Web Crawler*, navega entre as páginas HTML do banco de redações UOL de forma metódica e automatizada indexando textos de redações que posteriomente serão filtrados e coletados.

Na etapa subsequente a Figura 2 ilustra a normalização dos textos, que consiste em uma técnica de remoção de caracteres não alfa-numéricos presentes no HTML e espaços desnecessários, tal que o valor textual ainda seja o mesmo que o original. Após a normalização será organizado as as diversas partes que compõem a redação (tema, título, texto e nota) em uma estrutura JSON para armazenamento e uso futuro.

Figura 2: Os textos são submetidos aos algoritmos de normalização e posteriormente estruturados e armazenados no padrão JSON.

Na terceira etapa ilustrada pela Figura 3 será utilizada a ferramenta de mineração de dados *Orange* (LJUBLJANA, 2017). Será necessário realizar estudo e análise para obter o conhecimeto necessário para desenvolvimento de um fluxo de trabalho, seleção e treinamento dos modelos classificadores, concluindo todos os objetivos propostos nesta etapa.

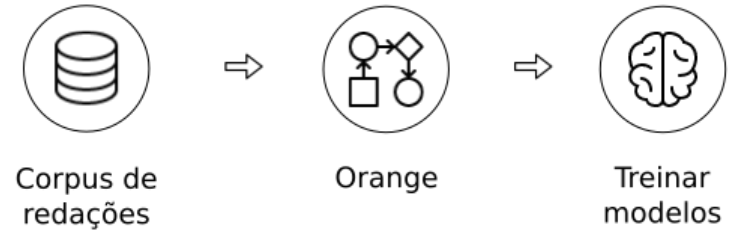

Figura 3: O *corpus* será utilizado em um fluxo de trabalho da ferramenta *Orange* para treinar os modelos classificadores.

A quarta e última etapa é ilustrada pela Figura 4, onde os modelos classificadores previamente ajustados e treinados serão submetidos aos testes de Acurácia (taxa de predições corretas ou incorretas realizada pelo modelo para um determinado conjunto de dados), *Overfitting* (super-ajustamento que ocorre quando o modelo se especializa nos dados utilizados no seu treinamento) e *Noise* (*noise* ou ruido é classificação errada do conjunto de dados de entrada), os resultados serão representados e comparados graficamente.

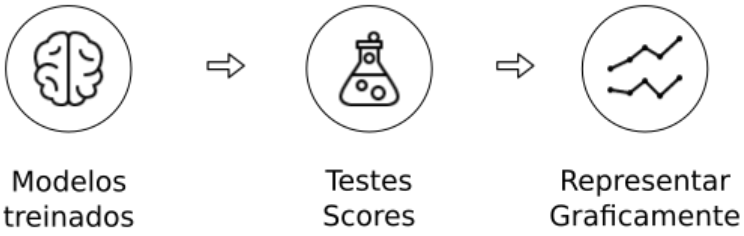

Figura 4: O modelos ajustados e treinados serão submetidos a testes, e os resultados comparados graficamente.

CAPÍTULO 5

DESENVOLVIMENTO

CAPÍTULO 6

RESULTADOS EXPERIMENTAIS

CAPÍTULO 7

CONCLUSÃO E TRABALHOS FUTUROS

Referências Bibliográficas

BATISTA, GEAPA; CARVALHO, ACPLF; MONARD, Maria C; BRASIL, Silicon Graphics. Aplicando seleção unilateral em conjuntos de exemplos desbalanceados: Resultados iniciais. In: **XIX CONGRESSO NACIONAL DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO "EDUCAÇÃO E APRENDIZAGEM NA SOCIEDADE DA INFORMAÇÃO**. [S.l.: s.n.], 1999. v. 20, p. 327–340.

CEBRASPE. **CENTRO BRASILEIRO DE PESQUISA EM AVALIAÇÃO E SELEÇÃO E DE PROMOÇÃO DE EVENTOS (CEBRASPE) PROGRAMA DE ATUALIZAÇÃO, QUALIFICAÇÃO E SELEÇÃO DE AVALIADORES DAS REDAÇÕES DO ENEM 2016** . 2016. Online; acessado 07 Abril 2017. Disponível em: <http://www.cespe.unb.br/colaboradores/ENEM_16_AVALIADOR_REDACAO/arquivos/ENEM_REDA____ES_2016___AVALIADOR___REGULAMENTO.PDF>.

CEBRASPE, CESPE UNB. **Relatório de Gestão CEBRASPE**. 2016. 1–20 p. Online; acessado 07 Abril 2017. Disponível em: <http://www.cespe.unb.br/cebraspe/arquivos/Relatorio_de_Gestao_2016.pdf>.

CSF, Ciência sem Fronteiras. **Estudante de Graduação**. 2017. Online; acessado 07 Abril 2017. Disponível em: <http://www.cienciasemfronteiras.gov.br/web/csf/estudante>.

INEP. **EDITAIS**. 2016. Online; acessado 07 Abril 2017. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/edital/2016/edital_enem_2016.pdf>.

LJUBLJANA, Bioinformatics Lab at University of. **Orange: Data Mining Toolbox in Python**. 2017. Online; acessado 01 Junho 2017. Disponível em: <https://orange.biolab.si/>.

MITCHELL, Tom M. Mitchell. **Machine Learning**. [S.l.]: Book News, Inc., 1997. ISBN 0070428077.

SISU, Sistema de seleção unificada. **O que é o Sisu**. 2017. Online; acessado 07 Abril 2017. Disponível em: <http://sisu.mec.gov.br/>.

UOL. **Banco de redações**. 2017. Online; acessado 01 Junho 2017. Disponível em: <https://educacao.uol.com.br/bancoderedacoes/>.

WAYMO. **We're building a safer driver for everyone**. 2017. Online; acessado 07 Abril 2017. Disponível em: <https://waymo.com/>.

APÊNDICE A – Título do Apêndice

APÊNDICE B – Exemplo do pacote Algorithm

Algoritmo 1 Estimador ML otimizado.

Inicializar o contador: $j \leftarrow 1$;
Fixar o limiar de variação das estimativas: $e_{\text{out}} \leftarrow 10^{-4}$;
Fixar o número máximo de iterações: $N \leftarrow 1000$;
Computar o ponto inicial: $\hat{\gamma}(0)$;
Determinar o limiar inicial: $e_1 \leftarrow 1000$;
Estabelecer o valor inicial de α: $\hat{\alpha}(0) \leftarrow -10^{-6}$;
enquanto $e_j \geq e_{\text{out}}$ e $j \leq M$ **fazer**
    Solucionar $\hat{\alpha}_j \leftarrow \arg\max_{\alpha} l_1(\alpha; \gamma_{j-1}, \mathbf{z}, n)$;
    Solucionar $\hat{\gamma}_j \leftarrow \arg\max_{\gamma} l_2(\gamma; \alpha_j, \mathbf{z}, n)$;
    $j \leftarrow j + 1$
    Computar o critério de convergência: e_j;
fim enquanto