

UNIVERSIDADE DE ITAÚNA



Recuperação de Padrões na Valoração Textual de Redações

Graduando: Eugênio Cunha

Orientador: Prof. Dr. Marco Túlio Alves N Rodrigues

27 de Novembro de 2017

Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Recuperação de Padrões na Valoração Textual de Redações

1. Introdução

Padrões

Aprendizado de Máquina

Problema de Pesquisa

2. Trabalhos Relacionados

3. Metodologia

4. Resultados Experimentais

5. Conclusões

6. Trabalhos Futuros

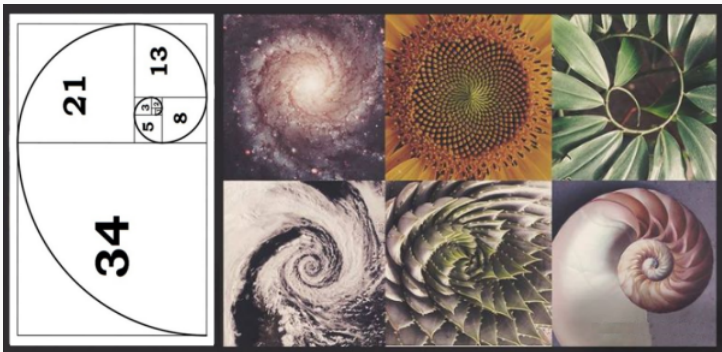
Introdução

Definição de Padrões



São perceptíveis **regularidades** que **repetem-se** de maneira **previsível** no mundo ou em um artefato produzido pelo homem.

Os Padrões no Mundo



Mendes (2007) explica, em seu estudo sobre a matemática na **natureza**, a ocorrência da **sequência** de **Fibonacci** na Natureza é tão frequente que é difícil acreditar que é acidental [6].

Os Padrões nos Artefatos






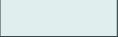



Ribeiro, L. (2007) em seu trabalho, classifica os estilos de pinturas rupestres do norte mineiro e sudoeste baiano [8].






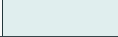



Souto, Lorena, Delbem e Carvalho explicam, o Aprendizado de Máquina, provê **técnicas** capazes de **aprender automaticamente** a partir dos **dados** disponíveis e produzir **hipóteses** úteis [4].

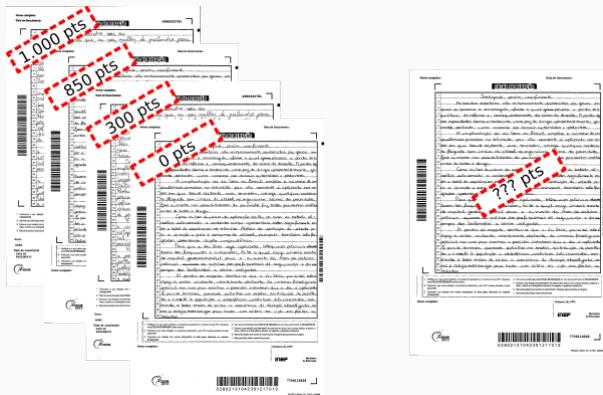
Qual é a **classe** da sétima linha?

| | Cores | R | G | B | Classe |
|---|---|-----|-----|-----|--------|
| 1 |  | 255 | 0 | 0 | Quente |
| 2 |  | 0 | 0 | 255 | Quente |
| 3 |  | 0 | 255 | 0 | Quente |
| 4 |  | 250 | 235 | 215 | Fria |
| 5 |  | 238 | 238 | 224 | Fria |
| 6 |  | 224 | 238 | 238 | Fria |
| 7 |  | 139 | 34 | 82 | ??? |

Encontrar **padrões**, **generalizar** e **predizer**!

| | Cores | R | G | B | Soma(RGB) | Classe |
|---|---|-----|-----|-----|-----------|---------------|
| 1 |  | 255 | 0 | 0 | 255 | Quente |
| 2 |  | 0 | 0 | 255 | 255 | Quente |
| 3 |  | 0 | 255 | 0 | 255 | Quente |
| 4 |  | 250 | 235 | 215 | 700 | Fria |
| 5 |  | 238 | 238 | 224 | 700 | Fria |
| 6 |  | 224 | 238 | 238 | 700 | Fria |
| 7 |  | 139 | 34 | 82 | 255 | Quente |

Problema de Pesquisa



Dado um *corpus* de redações rotuladas, é possível recuperar padrões implícitos nos textos e **valorar** uma nova amostra?

Trabalhos Relacionados

Matrizes de Referência

Silva (2017) explica em seu estudo, a prova de redação do ENEM é avaliada levando em conta uma matriz de referência elaborada pelo INEP [2].

| | Descrição | Valor |
|------------|---|-------|
| I | Demonstrar domínio da norma padrão da língua escrita. | 200 |
| II | Compreender a proposta de redação e aplicar conceitos das, várias áreas de conhecimento para desenvolver o tema, dentro,dos limites estruturais do texto dissertativo-argumentativo em prosa. | 200 |
| III | Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista. | 200 |
| IV | Demonstrar conhecimento dos mecanismos linguísticos, necessários para a construção da argumentação. | 200 |
| V | Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos. | 200 |

Matrizes de Referência

| III – Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista. | | |
|--|--|--------------|
| Nível | Descrição | Valor |
| 1 | Apresenta informações, fatos e opiniões relacionados ao tema proposto, de forma consistente e organizada, configurando autoria, em defesa de um ponto de vista. | <i>200</i> |
| 2 | Apresenta informações, fatos e opiniões relacionados ao tema, de forma organizada, com indícios de autoria, em defesa de um ponto de vista. | <i>150</i> |
| 3 | Apresenta informações, fatos e opiniões relacionados ao tema, limitados aos argumentos dos textos motivadores e pouco organizados, em defesa de um ponto de vista. | <i>100</i> |
| 4 | Apresenta informações, fatos e opiniões pouco relacionados ao tema ou incoerentes e sem defesa de um ponto de vista. | <i>50</i> |
| 5 | Apresenta informações, fatos e opiniões não relacionados ao tema e sem defesa de um ponto de vista. | <i>0</i> |

Bag-of-words

Nome completo: _____ Data de Nascimento: _____

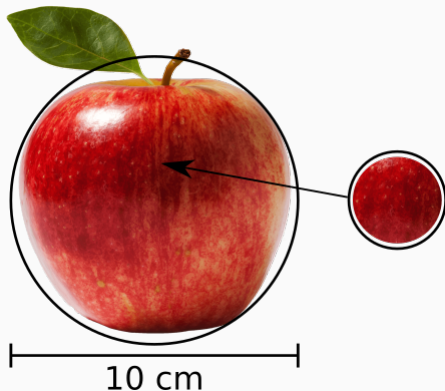
FOLHA DE RESPOSTAS

1. O primeiro parágrafo do texto apresenta uma visão geral da situação da saúde pública no Brasil, destacando a importância da prevenção e a necessidade de ações coordenadas entre os setores público e privado. A segunda parte do texto aborda a questão da distribuição de recursos, apontando para a necessidade de maior equidade na alocação de verbas, especialmente em relação ao acesso a serviços de saúde de qualidade. O terceiro parágrafo trata da importância da educação em saúde, destacando o papel das escolas e da mídia na promoção de hábitos saudáveis. O quarto parágrafo aborda a questão da vigilância epidemiológica, destacando a importância de sistemas de coleta e análise de dados para a identificação e controle de surtos. O quinto parágrafo trata da importância da pesquisa científica, destacando o papel das universidades e dos centros de pesquisa na geração de conhecimento para a melhoria da saúde pública. O sexto parágrafo aborda a questão da formação de recursos humanos, destacando a importância de cursos de graduação e pós-graduação na área da saúde. O sétimo parágrafo trata da importância da avaliação de programas, destacando a necessidade de métodos rigorosos para medir o impacto das intervenções. O oitavo parágrafo aborda a questão da transparência, destacando a importância de sistemas de prestação de contas para a população. O nono parágrafo trata da importância da participação social, destacando o papel da comunidade na tomada de decisões sobre a saúde pública. O décimo parágrafo aborda a questão da sustentabilidade, destacando a importância de modelos de financiamento que garantam a continuidade das ações de saúde pública.

| Índice | Frequência |
|--------|------------|
| 0 | 3.40 |
| 1 | 8.67 |
| 2 | 2.5 |
| 3 | 1.3 |
| 4 | 2.4 |
| 5 | 3.4 |
| ... | ... |
| 99 | 5.1 |
| 100 | 7.8 |

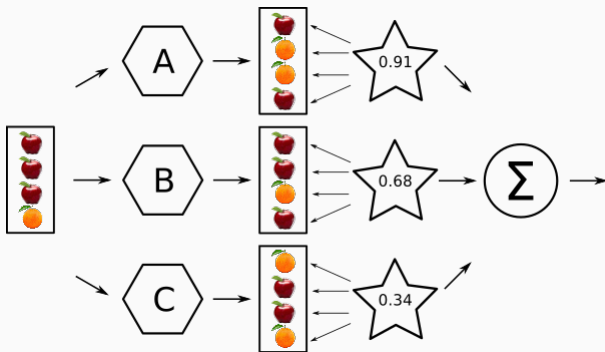
Matsubara, Martins e Monard (2003) explicam, um dos métodos adotados para simplificar a representação textual, é a abordagem *bag-of-words* [5].

Naive Bayes



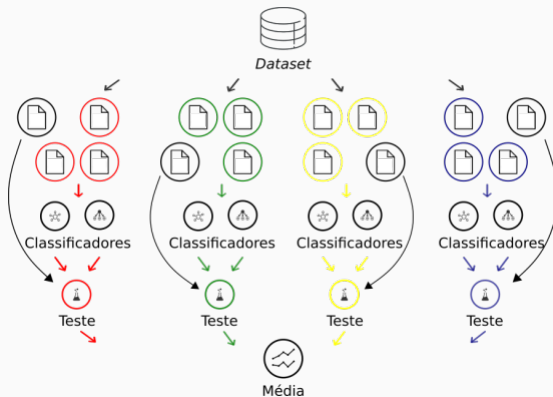
Brito (2017) descreve em sua pesquisa, o classificador Naive Bayes como um progenitor probabilístico [3].

Adaboost



Pereira e Ferreira (2015) explicam em sua pesquisa, Adaboost utiliza uma técnica que seleciona diversos algoritmos denominados classificadores fracos, com a finalidade de constituir um classificador forte [7].

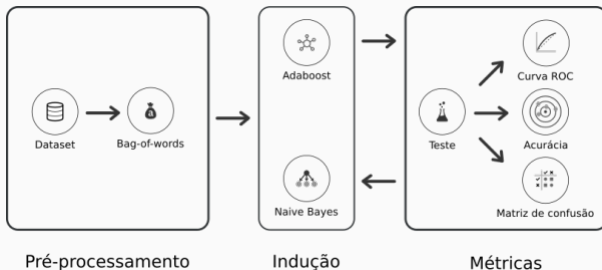
Validação cruzada



Baker, Isotani e Carvalho (2011) em seu trabalho demonstram, a validação cruzada permite verificar a corretude de um modelo gerado a partir da análise de dados de treinamento [1].

Metodologia

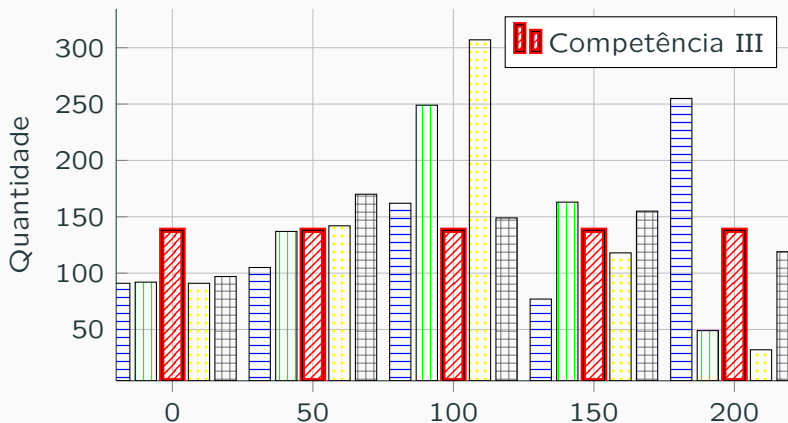
Inferência Indutiva



O *dataset* é submetido a técnica *bag-of-words* no pré-processamento, a estrutura de atributo-valor derivada é utilizada no treinamento dos classificadores, por fim, os classificadores são avaliados por métricas de desempenho.

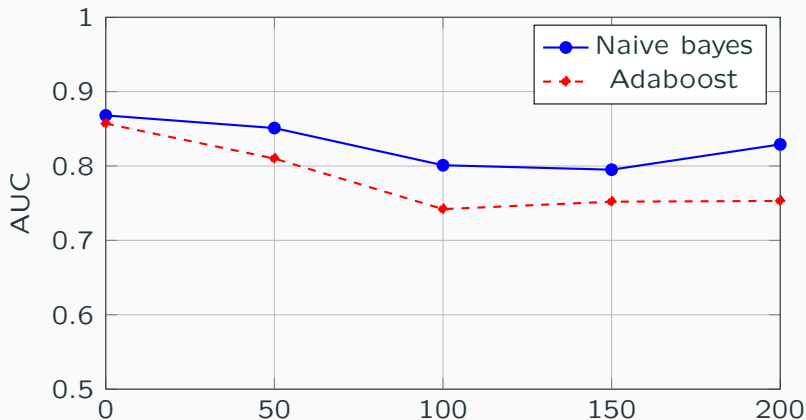
Resultados Experimentais

Disposição do Dataset



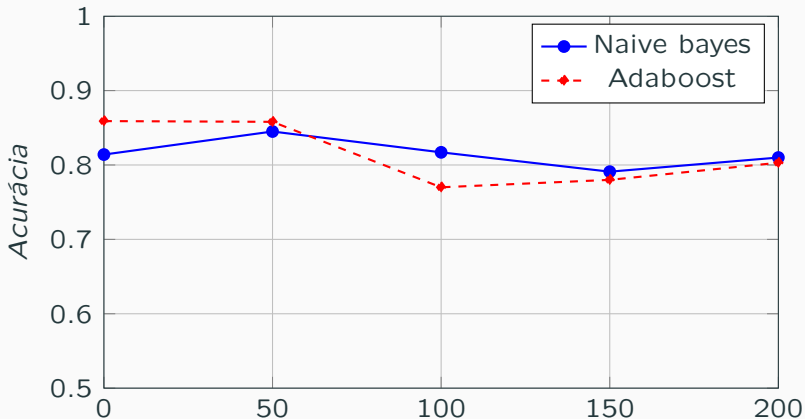
Distribuição das classes sobre a **competência III** de **690 redações** no *dataset* balanceado. Cada classe da competência III possui uma amostragem de **138 textos**.

Área Sobre a Curva ROC



O pontos da área sob curva ROC demonstra o **poder discriminativo** dos classificadores para cada uma das **classes** da competência III.

Acurácia



Os pontos da *acurácia* demonstram a **taxa de acerto**, comprovam que os classificadores **utilizam corretamente o poder de discriminação** de cada classe da competência III, para a rotulagem de novas amostras.

Matriz de Confusão

| | | <i>Predição do Naive Bayes</i> | | | | | |
|--------------|----------|--------------------------------|-----|-----|-----|-----|----------|
| | | 0 | 50 | 100 | 150 | 200 | Σ |
| <i>Atual</i> | 0 | 63 | 44 | 12 | 12 | 7 | 138 |
| | 50 | 3 | 99 | 33 | 1 | 2 | 138 |
| | 100 | 1 | 26 | 82 | 15 | 14 | 138 |
| | 150 | 3 | 10 | 17 | 61 | 47 | 138 |
| | 200 | 5 | 13 | 6 | 40 | 74 | 138 |
| | Σ | 75 | 192 | 150 | 129 | 144 | 690 |

A matriz de confusão é uma importante ferramenta, na **avaliação dos resultados** das predições, facilita visualmente o entendimento e **reage aos efeitos de predições falsas**.

Matriz de Confusão

| | | Predição do Naive Bayes | | | | | |
|-------|----------|-------------------------|-----|-----|-----|-----|----------|
| | | 0 | 50 | 100 | 150 | 200 | Σ |
| Atual | 0 | 63 | 44 | 12 | 12 | 7 | 138 |
| | 50 | 3 | 99 | 33 | 1 | 2 | 138 |
| | 100 | 1 | 26 | 82 | 15 | 14 | 138 |
| | 150 | 3 | 10 | 17 | 61 | 47 | 138 |
| | 200 | 5 | 13 | 6 | 40 | 74 | 138 |
| | Σ | 75 | 192 | 150 | 129 | 144 | 690 |

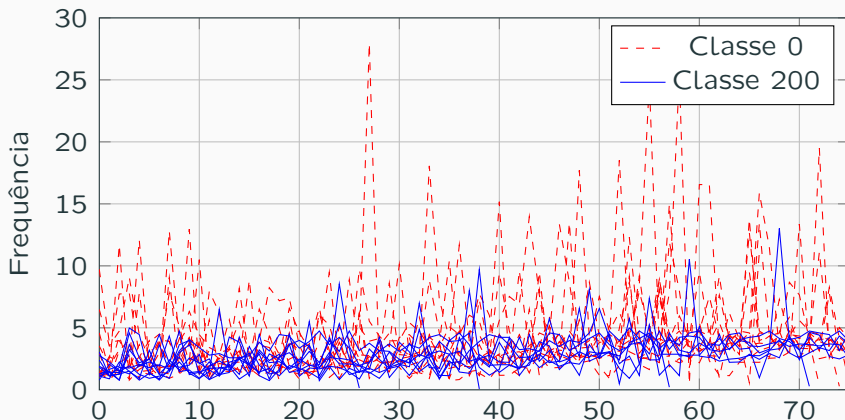
A matriz de confusão é uma importante ferramenta, na **avaliação dos resultados** das predições, facilita visualmente o entendimento e **reage aos efeitos de predições falsas**.

Matriz de Confusão

| | | Predição do Adaboost | | | | | |
|-------|----------|----------------------|-----|-----|-----|-----|----------|
| | | 0 | 50 | 100 | 150 | 200 | Σ |
| Atual | 0 | 99 | 18 | 10 | 9 | 2 | 138 |
| | 50 | 20 | 74 | 37 | 6 | 1 | 138 |
| | 100 | 8 | 27 | 75 | 20 | 8 | 138 |
| | 150 | 9 | 1 | 18 | 66 | 44 | 138 |
| | 200 | 7 | 5 | 13 | 50 | 63 | 138 |
| | Σ | 143 | 125 | 153 | 151 | 118 | 690 |

Quanto mais **próximas** as classes, maior é o número de confusões preditas pelos classificadores e quanto mais **distantes** as classes, menor o número de **confusões**, ou seja, ambos os classificadores tendem a confundir mais as classes **0** e **50**, do que as classes **0** e **200**.

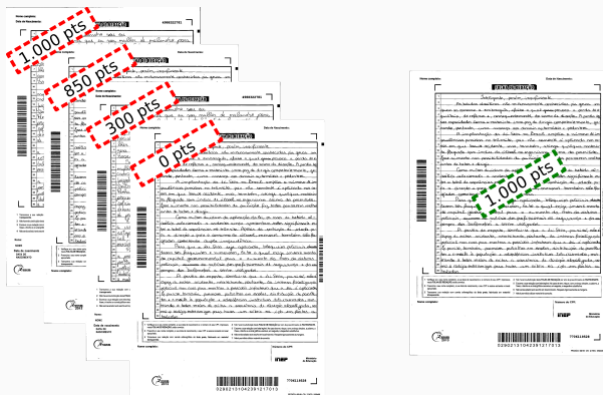
Padrões Recuperados



Ambos os padrões possuem **comportamentos similares**, entretanto, a **frequência** das palavras utilizadas **oscilam** em cada classe, o que demonstra a presença do padrão em cada redação.

Conclusões

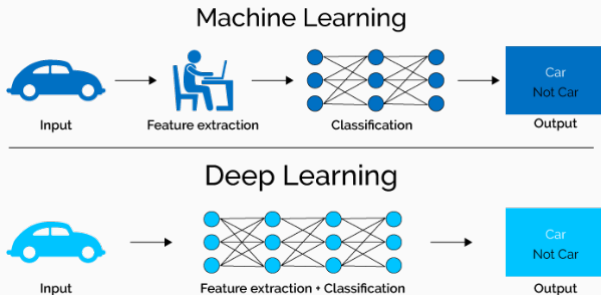
Conclusões



Os resultados obtidos, comprovam que os **classificadores** utilizados neste estudo, após a inferência indutiva, **retêm conhecimento**, para **recuperar padrões** e o **utilizam corretamente**, para a **predição** de novas amostras.

Trabalhos Futuros

Trabalhos Futuros



Em trabalhos futuros, pretende-se utilizar *Deep Learning*, para extrair um vetor numérico de características do texto, com o objetivo de mensurar com maior representatividade os padrões encontrados.

References I



R. Baker, S. Isotani, and A. Carvalho.

Mineração de dados educacionais: Oportunidades para o brasil.

Brazilian Journal of Computers in Education,
19(02):03, 2011.



S. R. da Silva and T. L. Carvalho.

**Produção de texto escrito no ensino médio:
Competências requeridas pela avaliação de
redação do enem em (des)uso no livro didático de
português.**

Caminhos em linguística aplicada, 16(1):1–25, 2017.

References II



E. M. N. DE BRITO.

Mineração de textos: detecção automática de sentimentos em co-mentários nas mídias sociais.

Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento, 6(1), 2017.



M. de Souto, A. Lorena, A. Delbem, and
A. de Carvalho.

Técnicas de aprendizado de máquina para problemas de biologia molecular.

Sociedade Brasileira de Computação, 2003.

References III



E. T. Matsubara, C. A. Martins, and M. C. Monard.
Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words.
Technical Report, 209, 2003.



F. M. P. Mendes.
A matemática na natureza.
Master's thesis, 2007.



M. U. Pereira and F. T. Ferreira.
Face detection.
2015.

References IV



L. Ribeiro.

Repensando a tradição: a variabilidade estilística na arte rupestre do período intermediário de representações no alto-médio rio são francisco.

Revista do Museu de Arqueologia e Etnologia,
(17):127–147, 2007.