

Classificação Automática de Textos por Período Literário Utilizando Compressão de Dados Através do PPM-C

Bruno Barufaldi¹, Eduardo F. Santana¹, José Rogério B. B. Filho¹, JanKees van der Poel², Milton Marques Júnior³, Leonardo Vidal Batista⁴

¹Departamento de Informática, Universidade Federal da Paraíba (UFPB)

²Programa de Pós-graduação em Engenharia Mecânica, Universidade Federal da Paraíba (UFPB)

³Departamento de Letras Clássicas Vernáculas, Universidade Federal da Paraíba (UFPB)

^{2,4}Programa de Pós-graduação em Informática, Universidade Federal da Paraíba (UFPB)

{bruno.barufaldi, eduardo.freire.87, jose.rogerio.filho}@gmail.com, jkvdpoel@yahoo.com.br, marquesjr45@hotmail.com

***Abstract.** Methods and techniques for data compression have been used for pattern recognition, including automatic text classification. The performance of the Prediction by Partial Matching (PPM) as a text classifier has already been proofed by many works, including authorship attribution for Portuguese texts. Classes involved in classification process may not be restricted by only one author. By including two or more authors in one class, one can create a literature style. This work presents a literature style classifier for texts from Brazilian literature by using the PPM-C statistical model.*

***Resumo.** Métodos e técnicas para compressão de dados têm sido utilizados para o reconhecimento de padrões, entre eles a classificação automática de textos. A eficiência do método Prediction by Partial Matching (PPM) como classificador textual já foi comprovada em diversos trabalhos, entre eles a atribuição de autoria para textos em português. As classes utilizadas no processo de classificação não precisam ficar restringidas a apenas um autor. Ao incluir dois ou mais autores numa mesma classe pode-se definir um estilo literário. Esse trabalho objetiva a construção de um classificador de períodos literários para textos da literatura brasileira utilizando o modelo estatístico PPM-C.*

1. Introdução

O aumento da popularidade da Internet nos últimos anos fez com que o número de dados circulando na rede crescesse abruptamente. Imagens digitais, textos e arquivos de áudio são armazenados e compartilhados entre usuários, muitas vezes com seu conteúdo marcado incorretamente e de forma não confiável. A maioria das ferramentas de busca na *World Wide Web* utiliza algoritmos para filtrar e detectar parâmetros textuais passados pelo usuário a fim de recuperar informação de forma automática, sem levar em consideração o conteúdo daquilo que se procura. Isso ocasiona um excesso de informações circulando atualmente na rede mundial que não conta com mecanismos inteligentes de busca ou classificação de conteúdo.

O Reconhecimento de Padrões pode ser definido como a disciplina da ciência cujo objetivo é a classificação de objetos em um determinado número de categorias ou classes [Theodoris e Koutroumbas, 2006]. Assim como os sinais da natureza estão sujeitos a regras e geram padrões, um texto – que pode ser entendido como um sinal – está sujeito a regras de linguagem e também gera padrões. Por esse motivo, o reconhecimento de padrões pode ser utilizado na Classificação Automática de Textos (CAT). As utilizações da CAT não se limitam em apenas melhorar mecanismos de busca, mas ela também pode ser utilizada em diversas aplicações, dentre elas a filtragem de *spam*, a identificação de conteúdo adulto, a organização de documentos em bibliotecas digitais e quaisquer outras aplicações que necessitem de seleção e organização de documentos.

O método de compressão de dados sem perdas *Prediction by Partial Match* (PPM) constrói um modelo estatístico a partir de uma determinada fonte de informação [Salomon, 2007]. Esse modelo é usado para diminuir a entropia dos símbolos da fonte e, assim, obter uma compressão sobre o sinal. Isso significa que quanto mais se conhece sobre a fonte, menor é a surpresa que seus símbolos causam ao aparecer e menor a quantidade de dados necessária para representá-los, resultando em uma maior compressão. Este método pode ser utilizado para o reconhecimento de padrões mapeando sinais (objetos) para modelos (classes) que obtiverem maior compressão sobre a fonte de informação [Coutinho et al., 2005].

A eficiência do PPM na classificação de textos já foi provada, superando inclusive classificadores Naïve Bayes, cujos modelos são baseados em palavras [Teahan e Harper, 2001]. Algumas vantagens da utilização de técnicas de compressão para classificação incluem a simplicidade e a ausência da necessidade de extração de características [Coutinho et al., 2005]. Atualmente, o PPM já está consolidado como um meio efetivo de atribuição de autoria para textos [Stamatatos, 2009][Coutinho et al., 2005].

Este trabalho tem como objetivo implementar um classificador automático de escolas literárias para textos da literatura brasileira utilizando o método de compressão de dados PPM-C. As escolas literárias Barroco, Arcadismo, Romantismo e Realismo foram contempladas no escopo deste trabalho.

2. Fundamentação Teórica

2.1. *Prediction by Partial Matching* (PPM) e Codificação Aritmética

A predição por emparelhamento parcial (*Prediction by Partial Matching*) é um dos mais eficientes métodos utilizados para compressão de dados sem perdas, sendo atualmente considerado o estado da arte nesta área. O PPM [Cleary e Witten, 1984] é um método para compressão de dados que mantém atualizado um modelo estatístico contextual adaptativo de uma fonte de informação [Salomon, 2007] que realiza a predição sobre o símbolo lido mais recente utilizando o seu contexto (o conjunto de símbolos precedentes), ao invés de trabalhar com as frequências referentes a cada símbolo da fonte de informação de forma isolada.

O modelo construído pelo PPM cria uma tabela para cada contexto utilizado em que a frequência de um determinado símbolo nesse contexto indica a probabilidade que

será utilizada na codificação. Neste trabalho foi utilizado o PPM-C [Moffat, 1990], uma das variantes do PPM.

O modelo PPM utiliza um conjunto de no máximo k símbolos precedentes como contexto para estimar a distribuição de probabilidades condicionais para o próximo símbolo da mensagem. Este modelo alimenta um codificador aritmético [Witten et al., 1987], que atribui a cada símbolo um número de bits inversamente proporcional à sua probabilidade condicional relativa ao contexto do símbolo.

Dado um novo símbolo S a ser comprimido em seu contexto C que é formado pelos símbolos precedentes e possui tamanho menor ou igual a k , o PPM utiliza seu modelo estatístico para calcular a probabilidade condicional da ocorrência do símbolo S no contexto C e passa essa probabilidade para o codificador aritmético. Caso o símbolo S não tenha ocorrido em um contexto de tamanho k , um símbolo de ESCAPE é codificado e é analisada a tabela que representa as probabilidades com contexto de tamanho $k-1$ (caso o símbolo não seja encontrado em nenhum dos contextos, ele é codificado utilizando um modelo que considera equiprováveis todos os símbolos possíveis). Após codificar o símbolo, o modelo atualiza as probabilidades condicionais do símbolo S . Este processo é repetido para cada novo símbolo a ser comprimido.

No final do processo, o codificador aritmético gera uma seqüência de símbolos codificados. Quanto menor for o tamanho dessa seqüência em relação ao tamanho do texto de entrada, maior será a compressão obtida.

A Tabela 1 mostra o modelo gerado pelo PPM após comprimir a cadeia de caracteres “hocuspocus”, utilizando um contexto com tamanho $k = 2$. “Contexto” indica os últimos símbolos lidos, c o contador desse símbolo (número de vezes que o símbolo apareceu num determinado contexto) e p sua probabilidade, derivada do seu contador.

Tabela 1. Modelo PPM depois do processamento da cadeia de caracteres *hocuspocus*.

Contexto $k = 2$				Contexto $k = 1$				Contexto $k = 0$		
Contexto	Símbolo	c	p	Predição	Símbolo	c	p	Predição	c	P
ho	c	1	$\frac{1}{2}$	h	o	1	$\frac{1}{2}$	h	1	$\frac{1}{16}$
	Esc	1	$\frac{1}{2}$		Esc	1	$\frac{1}{2}$	o	2	$\frac{2}{16}$
oc	u	2	$\frac{2}{3}$	o	c	2	$\frac{2}{3}$	c	2	$\frac{2}{16}$
	Esc	1	$\frac{1}{3}$		Esc	1	$\frac{1}{3}$	u	2	$\frac{2}{16}$
cu	s	2	$\frac{2}{3}$	c	u	2	$\frac{2}{3}$	s	2	$\frac{2}{16}$
	Esc	1	$\frac{1}{3}$		Esc	1	$\frac{1}{3}$	p	1	$\frac{1}{16}$
us	p	1	$\frac{1}{2}$	u	s	2	$\frac{2}{3}$	Esc	6	$\frac{6}{16}$
	Esc	1	$\frac{1}{2}$		Esc	1	$\frac{1}{3}$			
sp	o	1	$\frac{1}{2}$	s	p	1	$\frac{1}{2}$			
	Esc	1	$\frac{1}{2}$		Esc	1	$\frac{1}{2}$			
po	c	1	$\frac{1}{2}$	p	o	1	$\frac{1}{2}$			
	Esc	1	$\frac{1}{2}$		Esc	1	$\frac{1}{2}$			

Em geral, o PPM utiliza a codificação aritmética. Nela, a mensagem é representada inicialmente dentro do intervalo real $[0,1)$. Este intervalo é alterado à medida que os símbolos e suas probabilidades são inseridos no codificador. Quanto maior o tamanho da mensagem, menor o intervalo e mais casas decimais são necessárias para sua representação [Witten et al., 1987].

2.2. Literatura Brasileira

Literatura é a arte da palavra que atua como instrumento de comunicação e de interação social. Suas primeiras manifestações no Brasil ocorreram durante o período colonial (de

1500 a 1822), fortemente influenciada pela cultura portuguesa, tendo principalmente o propósito informativo. Atualmente, os poetas e prosadores se expressam de maneira diversificada, contribuindo com a arte mesmo sem que haja um projeto literário em comum [Cereja e Magalhães, 2002]. Apesar da origem da literatura brasileira ser bastante recente, comparada a outros países, a produção de textos literários no Brasil merece destaque e reconhecimento.

Um estilo literário pode ser entendido como um conjunto de textos com diversas características em comum. Apesar de não serem classificados como um mesmo estilo literário, o Barroco e o Arcadismo no Brasil são encontrados numa época, conhecida como fase luso-brasileira. Houve ecos do Barroco europeu entre os séculos XVII e XVIII, e sua transição para o Arcadismo buscou por esquemas rítmicos mais graciosos de forma específica e de menor beleza [Bosi, 2007]. Algumas características podem ser ressaltadas, tais como o cultismo e o conceptismo no Barroco, e o bucolismo e a simplicidade no conteúdo do Arcadismo.

Os períodos literários do Realismo e do Romantismo consolidaram-se no país e tiveram a contribuição de textos de diversos autores consagrados. Um dos traços essenciais do Romantismo brasileiro é o nacionalismo, explorando características como o indianismo, o regionalismo e a pesquisa histórica. Já os escritores realistas são motivados pelas teorias científicas e filosóficas da época, desejando retratar o homem e a sociedade em sua totalidade [Cereja e Magalhães, 2002].

3. Materiais e Métodos

Para classificar os textos, foram utilizadas quatro classes, as quais correspondem aos períodos literários Barroco, Arcadismo, Romantismo e Realismo. Os textos escolhidos estão listados a seguir, juntamente com seus respectivos autores e períodos literários, e foram inicialmente separados entre estas classes. Isto foi feito para tornar possível a posterior construção dos modelos estatísticos.

- **Barroco:** Antonio Vieira (*Sermão da Primeira Domingo do Advento, Sermão da Sexagésima, Sermão do Espírito Santo e Sermão do Bom Ladrão*) e Gregório de Matos (*Coletânea de Obras Líricas, Coletânea de Obras Satíricas e Coletânea de Obras Religiosas*);
- **Arcadismo:** Alvarenga Peixoto (*Coletânea de Obras*), Cláudio Manoel da Costa (*Poemas Escolhidos*), Basílio da Gama (*O Uruguai*) e Tomás Antônio Gonzaga (*Cartas Chilenas, Marília de Dirceu*);
- **Romantismo:** Joaquim Manuel de Macedo (*O Moço Loiro, A Moreninha, Os Dois Amores*), José de Alencar (*O Guarani, Senhora, Ubirajara, Iracema*), Machado de Assis (*A Mão e a Luva, Helena, Iaiá Garcia*), Manuel Antônio de Almeida (*Memórias de um Sargento de Milícias*) e Bernardo Guimarães (*A Escrava Isaura*);
- **Realismo:** Adolfo Caminha (*O Bom Crioulo, A Normalista*), Aluísio Azevedo (*O Mulato, O Homem, O Coruja*), Franklin Távora (*O Cabeleira*), Júlio Ribeiro (*A Carne*), Machado de Assis (*Memórias Póstumas de Brás Cubas, Dom Casmurro*) e Raul Pompéia (*O Ateneu, 14 de Julho na Roça, As Jóias da Coroa, Uma Tragédia no Amazonas*).

A coletânea de textos de Alvarenga Peixoto foi feita a partir dos poemas presentes no livro “A poesia dos inconfidentes: poesias completas de Cláudio Manuel da Costa, Tomás Antônio Gonzaga e Alvarenga Peixoto” de Domicio Proença Filho. As coletâneas de Gregório de Matos foram obtidas do livro “Poemas escolhidos: Gregório de Mattos” de José Miguel Wisnik [Mattos, 1999]. O restante dos textos foi obtido através do sítio Domínio Público [Portal Domínio Público, 2009] e do sítio Biblioteca Digital de Literatura do NUPILL [Biblioteca Digital de Literatura, 2009].

O processo de classificação pode ser dividido em três etapas: formatação dos textos, construção dos modelos e comparação da razão de compressão.

3.1. Formatação dos Textos

Antes da elaboração dos modelos e classificação, os textos passam por uma fase de padronização. São eliminados a acentuação e algumas pontuações e formatação, restando apenas as 26 letras do alfabeto (minúsculas) e os caracteres de espaçamento e ponto. Esta etapa tem por finalidade descartar símbolos pouco relevantes ou mesmo que dificultem a classificação correta, enquanto preserva a essência do texto, as palavras e frases.

3.2. Construção dos Modelos

Modelos são informações estatísticas sobre a ocorrência de símbolos dentro de determinados contextos que serão utilizadas para compressão. Uma vez criado, o modelo usado para classificação não será alterado.

Para cada texto a ser classificado, são gerados quatro modelos, sendo um para cada classe. Os modelos são construídos utilizando todos os textos presentes em uma classe. São lidos os n primeiros símbolos de cada texto, onde n é determinado pela razão entre o tamanho do treinamento e o número de textos por classe. O tamanho do treinamento é a quantidade de informação que será lida para construção do modelo, independentemente de quantos textos existam em uma determinada classe. Por exemplo, para um treinamento de tamanho 96kb e uma classe com oito textos, serão lidos os 12kb iniciais de cada texto. Essa informação é inserida no modelo desta classe.

Se um texto que se quer classificar pertença ao grupo de textos de uma classe, este texto não será utilizado na construção do modelo. Isto é feito para que não haja *vício* no treinamento: poderia ocorrer de um texto ser classificado de forma correta não por afinidade com os outros textos da classe, mas pela afinidade com ele próprio, o que não reflete a proposta do trabalho.

Os testes realizados utilizaram tamanhos de treinamento indo desde 8kb, passando por 16kb, 48kb, 96kb e chegando a 128kb.

3.3. Comparação da Razão de Compressão

O texto a ser classificado é comprimido por cada um dos quatro modelos segundo as regras do PPM-C. O texto será classificado como pertencente à classe cujo modelo obtiver maior compressão. Para classificação, foram feitos testes variando entre 0 e 10 os tamanhos máximos de contexto utilizados.

4. Resultados

Testes foram feitos variando tanto o tamanho dos modelos usados na classificação quanto o número de contextos utilizados pelo PPM. A Figura 1 mostra que a maior taxa de acerto médio foi de 85%, encontrada ao se utilizar 48kb para efetuar o treinamento dos modelos. Este índice médio de acertos representa a média de acertos encontrada em cada contexto testado, a saber, indo do contexto $k = 0$ até o contexto $k = 11$.

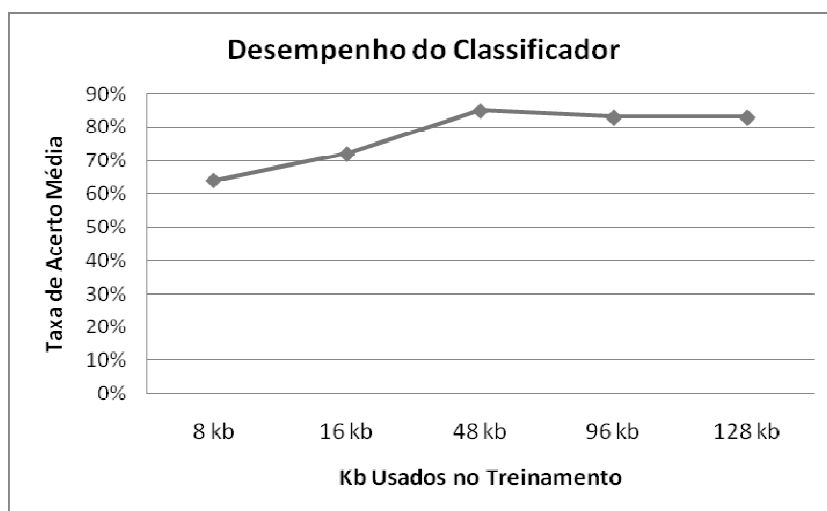


Figura 1. Gráfico de acerto médio por tamanho de treinamento.

A Figura 2 mostra os resultados obtidos separadamente em cada contexto utilizando 48 Kb para treinamento do classificador. Pode-se observar que o melhor resultado foi encontrado ao se utilizar contexto $k = 4$. Mesmo aumentando o número de contextos é notada uma pequena queda no desempenho do classificador. Tal fato acontece devido à natureza de compressão do PPM, cuja curva de aprendizado pára de crescer a partir de certo contexto ocasionando uma redução na compressão. [Salomon, 2007].

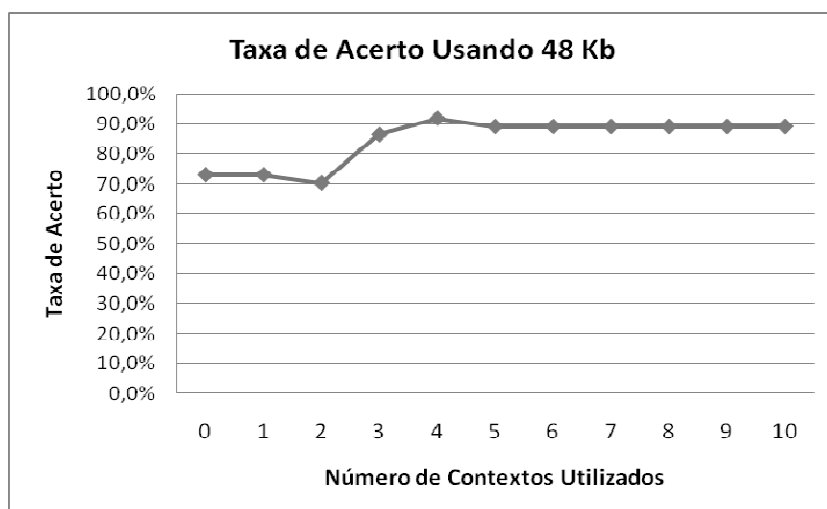


Figura 2. Gráfico de acerto obtido na classificação usando 48kb e diferentes contextos.

A Tabela 2 é a tabela de confusão obtida na classificação quando se usa 48 Kb para treinamento e um contexto $k = 4$. Dela pode-se inferir que apenas três textos foram classificados erroneamente: a coletânea de obras líricas de Gregório de Matos, Helena de Machado de Assis e Memórias de Um Sargento de Milícias de Manuel Antônio Bandeira. Possíveis razões para esse erro na classificação são discutidos na próxima sessão deste trabalho.

Tabela 2. Tabela de confusão da classificação usando 48 Kb e contexto $k = 4$.

Estilos Literários/Obras		Classificadas como	Barroco	Arcadismo	Romantismo	Realismo
Barroco	Sermão da Primeira Domingo do Advento		X			
	Sermão da Sexagésima		X			
	Sermão do Espírito Santo		X			
	Sermão do Bom Ladrão		X			
	Coletânea de obras líricas			X		
	Coletânea de obras satíricas		X			
	Coletânea de obras religiosas		X			
Arcadismo	Coletânea			X		
	Poemas Escolhidos			X		
	O Uruguai			X		
	Cartas Chilenas			X		
	Marília de Dirceu			X		
Romantismo	O Moço Loiro				X	
	A Moreninha				X	
	Os Dois Amores				X	
	O Guarani				X	
	Senhora				X	
	Ubirajara				X	
	Iracema				X	
	A Mão e a Luva				X	
	Helena					X
	Iaiá Garcia				X	
	Memórias de um Sargento de Milícias					X
	A Escrava Isaura				X	
Realismo	O Bom Crioulo					X
	A Normalista					X
	O Mulato					X
	O Homem					X
	O Coruja					X
	O Cabeleira					X
	A Carne					X
	Memórias Póstumas de Brás Cubas					X
	Dom Casmurro					X
	O Ateneu					X
	14 de Julho na Roça					X
	As Jóias da Coroa					X
	Uma Tragédia no Amazonas					X

5. Conclusões e Discussões

Obteve-se uma taxa de acerto máxima de 91,89% utilizando 48 Kb no treinamento e utilizando o PPM com número de contextos $k = 4$. Com esses parâmetros de treinamento e compressão, ocorreram apenas três classificações incorretas: Helena (Machado de Assis), Gregório de Matos no estilo lírico e Memórias de Um Sargento de Milícias. Estes erros podem ser atribuídos às particularidades presentes nessas obras.

Machado de Assis, romancista consagrado entre especialistas da área, tem características marcantes que iniciaram o movimento realista no país. Apesar de a classificação pelo PPM obter resultados satisfatórios em suas obras, o marcante “estilo machadiano” pode influenciar nos resultados, considerando que textos do autor foram utilizados tanto na construção do modelo romântico quanto na construção do modelo realista. Como exemplo disto, o romance “Helena” foi classificado como realista, um equívoco que não se repetiu em outros textos de sua autoria.

Gregório de Matos, um dos autores barrocos utilizados na pesquisa, possui características distintas dos outros autores. Suas obras foram selecionadas e associadas a estilos satíricos, líricos e religiosos. Contudo, Gregório de Matos no estilo lírico persiste na classificação árcade com uma diferença de compressão em torno de 3% para a compressão obtida pelo modelo barroco. A utilização de referentes clássicos e algumas metáforas com elementos da natureza nos textos líricos podem ter influenciado sua classificação como árcade.

A classificação da obra “Memórias de um Sargento de Milícias”, romance de Manuel Antônio de Almeida, como sendo de estilo Realista pode ser justificada pela mesma pertencer à transição entre os estilos Romântico e Realista. Apesar desta sua obra ser do início do Romantismo, a mesma possui características que antecipam o Realismo. Por possuir característica de ambos os estilos, ele foi muitas vezes classificado como realista. Contudo, em todas as classificações incorretas o modelo romântico conseguiu obter a segunda melhor compressão, com uma diferença de 1% para o modelo realista.

Uma das maiores dificuldades encontradas durante a pesquisa que resultou no presente artigo foi a pouca disponibilidade de textos originais no formato digital. Sendo assim, atualmente a pesquisa está focada em obras barrocas, árcades, românticas e realistas. Apesar disto, tem-se a perspectiva de refinar o modelo criado através da inserção de novos textos e estilos literários, o que levaria a uma melhora dos resultados.

Trabalhos futuros irão estudar a utilização de atributos textuais para auxiliar a classificação automática de textos em conjunto com o PPM. Esta abordagem investigaria uma possível melhora na classificação dos textos cujo estilo literário tido como correto ficou em segundo lugar nos resultados. Esses atributos seriam dispostos em um vetor de características, tais como tamanho médio das palavras, riqueza vocabular e entropia dos bigramas. Com esse vetor pode ser feita uma segunda classificação do texto em questão através de um método como a distância euclidiana mínima entre o vetor e o vetor representante de cada classe. Caso a diferença entre a razão de compressão dos dois primeiros lugares estiver dentro de um limiar a definir seria utilizada a classificação por atributos textuais.

Cabe aqui salientar que não existem na literatura pesquisas utilizando o PPM (ou quaisquer outros métodos) para classificar textos por *período literário* (ou até mesmo que tratem de problemas similares). Por conta deste fato, não é possível efetuar maiores comparações.

O Professor Milton Marques Junior, doutor em Letras pela Universidade Federal da Paraíba, auxiliou na pesquisa que culminou com o presente artigo, colaborando com seus conhecimentos na área. Por ser um especialista, o professor orientou os alunos através da disponibilização de textos e discussões relacionadas à literatura brasileira.

Referências

Biblioteca Digital de Literatura. Núcleo de Pesquisas em Informática, Literatura e Linguística da UFSC (NUPILL). Disponível em <<http://www.literaturabrasileira.ufsc.br/>>. Acessado em 24 de maio de 2009.

Bosi, A. (2007). “História concisa da Literatura Brasileira”, 44ª Edição.

Cereja, W. R.; Magalhães, T. C. (2002). “Literatura Brasileira”, 2ª Edição.

- Cleary, J.G.; Witten, I. H. (1984). "Data compression using adaptive coding and partial string matching", IEEE Transactions on Communications, v. 32, n. 4, pp. 396-402.
- Coutinho, B. C.; Macedo, J. L. de M.; Júnior, A. R.; Batista, L. V. (2005). "Atribuição de Autoria usando PPM". In: III Workshop em Tecnologia da Informação e da Linguagem Humana, 2005, São Leopoldo. Anais do XXV Congresso da Sociedade Brasileira de Computação, 2005. v. 1. p. 2208-2217.
- Mattos, G. (1999). "Poemas escolhidos: Gregório de Mattos"; seleção, introdução e notas de José Miguel Wisnik. 7ª Edição. São Paulo: Cultrix.
- Moffat, A. (1990). "Implementing the PPM data compression scheme". IEEE Transactions on Communications, v. 38, n.11, pp. 1917-1921.
- Peixoto, A. (1996). "Poesias". In: "A poesia dos inconfidentes: poesias completas de Cláudio Manuel da Costa, Tomás Antônio Gonzaga e Alvarenga Peixoto"; organização de Domício Proença Filho; artigos, ensaios e notas de Eliana S. Muzzi, João Ribeiro, Letícia Malard, Lúcia Helena, Luciano Figueiredo, Manuel Bandeira, Manuel Rodrigues Lapa, Melânia Silva de Aguiar e Paulo Roberto Dias Pereira. Rio de Janeiro: Nova Aguilar.
- Portal Domínio Público. Disponível em <<http://www.dominiopublico.gov.br/>>. Acessado em 24 de maio de 2009.
- Salomon, D. (2007). Data Compression, 4th Edition.
- Stamatatos, E. (2009). "A survey of modern authorship attribution methods". Journal of the American Society for Information Science and Technology, v. 60, n. 3, pp. 538-556.
- Teahan, W. J.; Harper, D. J. (2003). "Using compression-based language models for text categorization". In: W. B. Croft and J. Lafferty (Eds.), Language Modeling for Information Retrieval, pp. 141-166. Kluwer Academic Publishers, 2003.
- Theodoris, S.; Koutroumbas, K. (2006), "Pattern Recognition", 3rd Edition.
- Witten, I. H.; Neal, R. M.; Cleary, J. G. (1987). "Arithmetic Coding For Data Compression". In Journal of the ACM, v. 30, n. 6.