Um Método para Coleta, Tratamento de uma Base de Dados de Textos Literários em Língua Portuguesa para uso na Identificação de Autoria

Paulo Júnior Varela[1], Jivago Bulsing Schoot[1]

[1]Coordenação de Licenciatura em Informática - COLIN
Universidade Tecnológica Federal do Paraná (UTFPR) – Francisco Beltrão, PR – Brasil
paulovarela@utfpr.edu.br, jivag0w@hotmail.com

Resumo. *A utilização do meio computacional para a resolução de casos de identificação de autoria tem crescido progressivamente em áreas como a computação, a linguística e o direito. Este artigo tem por objetivo apresentar um método para auxiliar no processo de coleta, tratamento e validação de base de dados para fins de testes de modelos computacionais. Como base de dados foram coletados, tratados e validados 100 amostras de textos de autores consagrados da literatura brasileira. Ao final a base de dados foi validada e com as características utilizadas tal base se mostrou robusta e confiável para utilização no processo de identificação de autoria de textos.*

Abstract. *The use of computational means for resolving cases of authorship identification has grown progressively in areas such as computing, linguistics and law. This article aims at presenting a method to aid in the process of collecting, processing and validation database for testing computational models. As database were collected, processed and validated 100 examples of works by renowned authors in Brazilian literature. At the end of the database has been validated and used with features such base proved robust and reliable for use in the process of authorship identification of texts.*

1. Introdução

A linguística forense é um ramo da linguística que estuda os diversos pontos de encontro entre a linguagem e a lei, com o fim de apontar evidências linguísticas nos processos judiciais. Pode-se denotar que as relações entre a linguística, o direito e a computação estão se aproximando com o passar dos tempos, como resultado do interesse de linguistas, juristas e pesquisadores que estão mostrando em suas respectivas áreas os resultados alcançados com suas aplicações práticas.

A análise estilística forense busca estabelecer um dicionário de atributos estilométricos, como parâmetro estável de análise das variabilidades entre escritores distintos, tais como: a frequência de palavras incomuns; a média do tamanho das orações; o quociente de palavras diferentes em relação ao total; entre outros. Portanto, é possível afirmar que o conjunto de valores obtidos pela quantização de tais atributos definirá o estilo [McMenamim, 2002]. O foco da estilistica forense é a identificação do autor em documentos cuja autoria seja questionada.

Na esfera jurídica, muitos processos estão inter-relacionados diretamente com o questionamento da autoria de documentos impressos e digitais. Podem se citar vários

exemplos que podem ser encontrados nesse meio, tais como: cartas de ameaça, cartas de sequestro, e-mails, notas de resgate, bilhetes e cartas de difamação, cartas de suicídios, livros, artigos e colunas em panfletos, jornais e revistas, bem como os demais documentos que cuja autoria seja desconhecida e a análise da grafia não seja possível de aplicar para identificar o autor [Pavelec *et al*., 2007][Varela *et al*., 2011].

Atualmente, a utilização de tais documentos como prova fica sujeita a análise por parte dos peritos designados pelos juízes. No entanto, este processo de análise ainda é pouco conhecido e utilizado no Brasil. O principal ponto crítico da análise destes documentos por partes dos peritos, é que os mesmos não possuem um método padrão de análise e nem mesmo ferramentas que possam auxiliar na identificação de autores de língua portuguesa. Cabe ressaltar as questões da imprecisão dos métodos linguísticos, que sofrem ainda com a influência demasiada do perito e de sua subjetividade [Varela, 2010].

Para auxiliar no processo de desenvolvimento e robustez de um modelo de identificação de autoria é necessário efetuar diversos testes a fim de validar o modelo, para que este possa posteriormente ser aplicado em um ambiente real [Johnstone, 2005]. Sendo assim, a coleta de uma base de dados para realização dos experimentos e validação se faz de suma importância para o modelo a ser desenvolvido.

O objetivo principal deste artigo é demonstrar de forma detalhada um protocolo de coleta e tratamento de uma base de dados para que esta possa ser utilizada em experimentos de identificação de autoria.

Este artigo apresenta algumas seções importantes para a compreensão do trabalho, que são: a metodologia de coleta e tratamento dos dados, que demonstra o processo de coleta, tratamento e extração das características dos textos. Após é detalhado como foi montado os vetores de características para uso no processo de classificação, e ao final um pequeno relato da validação que transformou o trabalho viável.

2. Metodologia de Coleta e Tratamento da Base de Dados

Para o desenvolvimento de uma base de dados que possa ser utilizada em testes de identificação é necessário primeiramente estabelecer alguns critérios para realizar uma boa coleta de dados. Sendo assim, primeiramente foi desenvolvido um protocolo de coleta de textos, apresentado na Figura 1. Neste protocolo, a primeira fase constou da definição das características que a base teria que conter (textos de capítulos ou parte de capítulo de obras consagradas da literatura brasileira que tivessem mais que 1000 palavras). Em um segundo momento, se procedeu a definição de quais autores e quais obras seriam coletadas para a base, bem como a quantidade de autores e obras. Logo em seguida aconteceu a garimpagem de sítios na internet que disponibilizassem obras da literatura brasileira em domínio público, sendo neste caso as bibliotecas virtuais desenvolvidas pelo governo federal e universidades as principais fontes de pesquisa. Depois de encontradas as obras foi efetuado o download das obras separando-as por autor e por período literário ao qual a obra pertence (realismo, romantismo, parnasianismo, modernismo ou contemporâneo). De posse dos textos, houve então o tratamento dos textos (mais detalhes na seção 2.2). E por fim, para validar a base foram realizados alguns testes para extração de características e classificação.

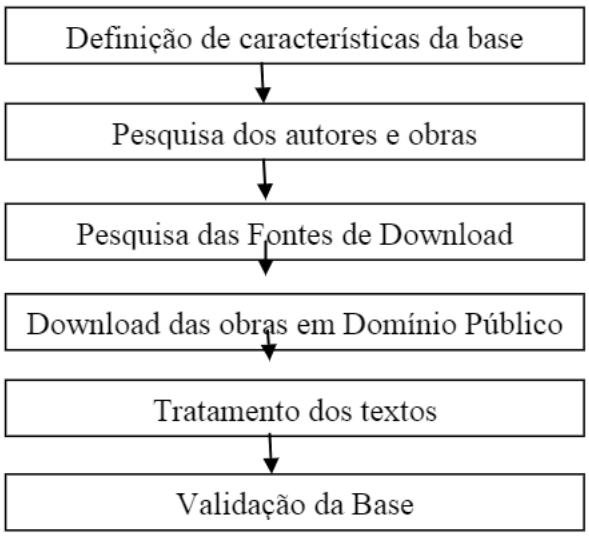


Figura 1 – Protocolo de Coleta e Tratamento da Base de Dados

2.1 Coleta da Base de Dados

A base de dados foi concebida por 25 autores consagrados da literatura brasileira. Foram coletadas obras literárias em domínio público e disponibilizadas via bibliotecas digitais na internet, tais como: Biblioteca Nacional (www.dominiopublico.gov.br), bibliotecas virtuais da Universidade de São Paulo, Universidade Federal de Santa Catarina e organizações não governamentais, tal como a Gutenberg.org.

Os textos dos 25 autores foram separados em 5 períodos literários aos quais as obras pertencem, que são: romantismo, realismo, modernismo, parnasianismo e contemporâneo. Os autores suas respectivas obras são relacionados na Tabela 1. Para cada autor foram coletadas 4 amostras de textos com mais de 1000 caracteres. Tais amostras são capítulos ou partes de capítulos das obras relacionadas.

As obras coletadas estavam em formato .pdf (*Portable Document Format*) e foram transformados para o formato de caracteres ASCII, ou seja, arquivo texto com extensão .txt (*Text File*) . Foram realizadas as devidas formatações para eliminação de textos acessórios, tais como: cabeçalhos, rodapés, número de páginas e "sujeiras textuais" que não constam na obra original.

Para organizar a base de dados foi utilizado como nomenclatura de identificação de cada amostra de texto, as iniciais do autor seguido do número da amostra. Por exemplo: As amostras do autor José de Alencar são identificadas na base como: JA01, JA02, JA03 e JA04.

Tabela 1 - Base de Textos (Autores e Obras do Romantismo)

ROMANTISMO	
AUTOR	**LIVRO**
JOSÉ DE ALENCAR	O GUARANI

VISCONDE DE TAUNAY	INOCÊNCIA
JOAQUIM MANUEL DE MACEDO	A LUNETA MÁGICA
FRANKLIN TÁVORA	O CABELEIRA
BERNARDO GUIMARÃES	A ESCRAVA ISAURA

Tabela 2 - Base de Textos (Autores e Obras do Realismo)

REALISMO	
AUTOR	**LIVRO**
MACHADO DE ASSIS	MEMÓRIAS PÓSTUMAS DE BRÁS CUBAS
RAUL POMPÉIA	O ATENEU
ADOLFO CAMINHA	A NORMALISTA
ALUISIO AZEVEDO	O CORTIÇO
INLGÊS DE SOUZA	O SEMINÁRIO

Tabela 3 - Base de Textos (Autores e Obras do Modernismo)

MODERNISMO	
AUTOR	**LIVRO**
JOÃO DO RIO	A ALMA ENCANTADA DAS RUAS
MÁRIO DE SÁ CARNEIRO	A CONFISSÃO DE LÚCIO
ALCANTARA MACHADO	BRAS, BEXIGA E BARRA FUNDA
JACKSON DE FIGUEIREDO	A REAÇÃO DO BOM SENSO
ALBERTO TORRES	AS FONTES DE VIDA DO BRASIL

Tabela 4 - Base de Textos (Autores e Obras do Parnasianismo)

PARNASIANISMO	
AUTOR	**LIVRO**
OLAVO BILAC	CONTOS PARA VELHOS
AMADEU AMARAL	MEMORIAL DE UM PASSAGEIRO DE BONDE
RUI BARBOSA	OBRAS SELETAS
PAULO SETUBAL	A MARQUESA DE SANTOS
AUGUSTO DOS ANJOS	EU E OUTRAS POESIAS

Tabela 5 - Base de Textos (Autores e Obras do Contemporâneo)

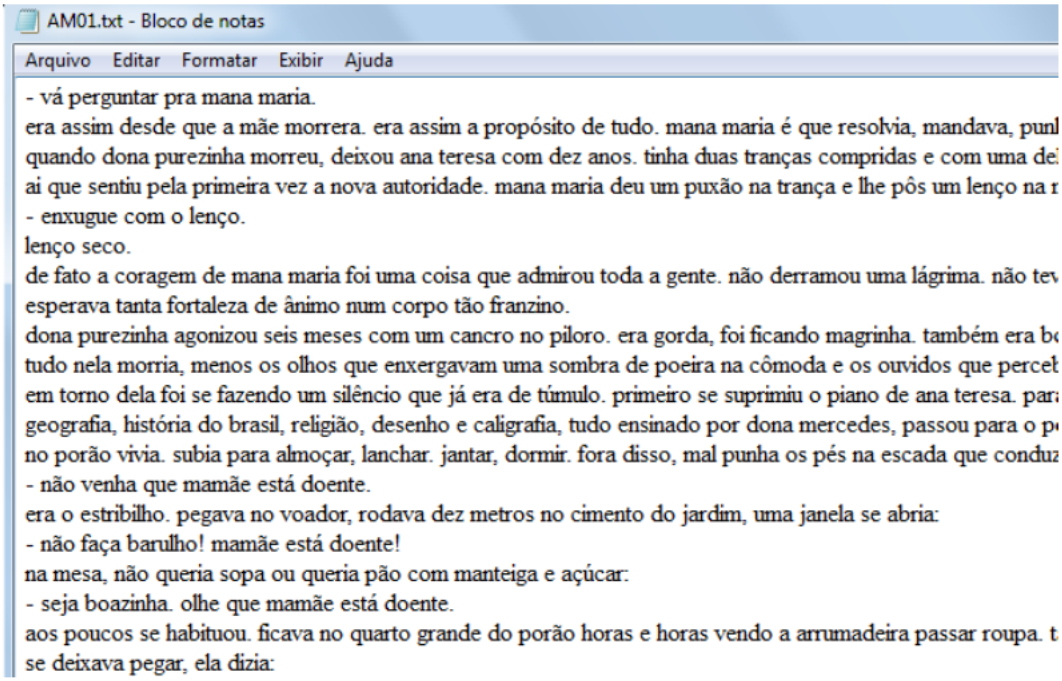
CONTEMPORÂNEO	
AUTOR	**LIVRO**
JOSÉ GUIMARÃES	A CAMPANHA

AFRÂNIO PEIXOTO	HISTÓRIA DO BRASIL
HUGO MÁXIMO	MUNDO BIZARRO
ERNESTO ROSA	O JOGO DO VARDIÃO
CEZAR DIAS	TUBARÃO COM A FACA NAS COSTAS

2.2 Processo de Tratamento da Base de Dados

Para que os textos da base de dados fossem homogêneos foi realizado o processo de tratamento, que significa a retirada de elementos textuais que possam prejudicar no processo de reconhecimento. Tais elementos removidos foram: número de páginas, informações de cabeçalho e rodapé e outras informações desnecessária, conhecidas como "sujeira textual".

Como principio fundamental para a extração de características, os textos foram padronizados em letras minúsculas, em formato ASCII, armazenados em arquivos de texto (txt). Esta padronização se faz necessária para que a ferramenta de extração de características consiga atingir o máximo de aproveitamento possível, pois pequenas características podem ser essenciais para o processo de reconhecimento de padrões em textos. Um exemplo de amostra de texto pode ser visto na Figura 2.


AM01.txt - Bloco de notas

Arquivo Editar Formatar Exibir Ajuda

- vá perguntar pra mana maria.
era assim desde que a mãe morrera. era assim a propósito de tudo. mana maria é que resolvia, mandava, punl
quando dona purezinha morreu, deixou ana teresa com dez anos. tinha duas tranças compridas e com uma del
ai que sentiu pela primeira vez a nova autoridade. mana maria deu um puxão na trança e lhe pôs um lenço na r
- enxugue com o lenço.
lenço seco.
de fato a coragem de mana maria foi uma coisa que admirou toda a gente. não derramou uma lágrima. não tev
esperava tanta fortaleza de ânimo num corpo tão franzino.
dona purezinha agonizou seis meses com um cancro no piloro. era gorda, foi ficando magrinha. também era bo
tudo nela morria, menos os olhos que enxergavam uma sombra de poeira na cômoda e os ouvidos que percel
em torno dela foi se fazendo um silêncio que já era de túmulo. primeiro se suprimiu o piano de ana teresa. para
geografia, história do brasil, religião, desenho e caligrafia, tudo ensinado por dona mercedes, passou para o po
no porão vivia. subia para almoçar, lanchar. jantar, dormir. fora disso, mal punha os pés na escada que conduz
- não venha que mamãe está doente.
era o estribilho. pegava no voador, rodava dez metros no cimento do jardim, uma janela se abria:
- não faça barulho! mamãe está doente!
na mesa, não queria sopa ou queria pão com manteiga e açúcar:
- seja boazinha. olhe que mamãe está doente.
aos poucos se habituou. ficava no quarto grande do porão horas e horas vendo a arrumadeira passar roupa. t
se deixava pegar, ela dizia:

Figura 2 – Amostra de texto da base de dados tratado

2.3 Extração de Características

O processo de extração de características contou com o auxílio do software Lexico 3[1], que foi utilizado para extração das características de cada amostra de texto. Após o processo de extração foi gerado um relatório com as informações estatísticas dos textos, que inclui: número de total de palavras, quantidade de formas de palavras, frequência máxima de uma palavra, número de *hapax legonema*[2].

Tabela 6 – Informações estatísticas dos Textos

Total de Palavras	Formas	Frequência	Hapax	Autor
1450	599	48	401	AM
1524	653	61	456	AM
1834	749	71	536	AM
2189	878	89	600	AM
1469	681	76	518	AT
2091	901	625	694	AT
3020	1000	132	676	AT
2703	894	117	604	AT

3. Geração dos Vetores de Características

O método proposto se baseia nos procedimentos de análise estilística forense, que estabelece a associação ou dissociação da autoria do texto, em relação a um provável autor ou um período literário, como base num conjunto de atributos estilométricos previamente estabelecido. A associação indica a existência de atributos estilométricos suficiente para garantir estatisticamente, que o texto, de autoria desconhecida, pertence ao período avaliado. A dissociação indica que o mesmo não pertence ao período avaliado.

Do ponto de vista computacional, o modelo é conhecido como modelo global, pois um único modelo é utilizando na associação ou dissociação da autoria do texto questionado.

Ao final é gerado um vetor de características de cada amostra de texto, que é utilizado posteriormente para a realização da classificação dos textos. Tais vetores são compostos a partir de informações da frequência das características utilizadas.

Os vetores gerados, constam de características e indicação de uma classe, como pode ser visto no exemplo da Figura 3. As primeiras posições do vetor indicam as informações de constantes no texto, e a última informação mostra a qual classe (período literário) o texto pertence.

[1] Software da Université de la Sorbonne Nouvelle - Paris http://www.tal.univ-paris3.fr/lexico/index-gb.htm

[2] Quantidade de tokens que não se repetem.

```
,0,30,0,6,2,4,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,2,1,0,6,0,3,4,AP
0,3,2,2,4,0,80,0,6,4,6,0,10,0,1,0,2,0,0,0,3,2,0,1,1,4,3,2,1,4,2,4,2,10,0,0,8,AP
,5,3,1,2,0,50,0,3,3,9,0,4,0,0,0,2,0,0,0,0,2,0,4,0,0,1,7,2,4,0,8,0,5,0,1,12,AP
0,1,2,0,41,0,4,3,5,0,9,0,1,0,1,0,0,0,2,0,0,5,0,0,3,4,0,1,0,3,0,7,0,0,4,AP
,19,0,0,0,0,0,2,0,0,0,0,0,0,1,0,0,1,0,0,0,0,3,2,0,0,1,0,0,0,2,0,CD
,14,0,0,0,1,0,4,0,0,0,0,0,0,3,1,1,0,1,0,0,0,0,0,0,0,1,0,0,1,0,1,1,CD
,15,0,1,0,2,0,5,0,0,0,1,0,4,0,0,1,1,0,0,0,1,1,1,0,0,1,0,5,0,2,0,CD
,7,0,1,3,2,0,0,0,0,0,0,0,0,3,0,0,0,0,0,1,0,0,0,0,0,2,0,0,0,0,0,0,0,CD
,18,0,1,1,2,0,4,0,0,0,1,0,1,1,0,2,0,1,0,0,3,4,0,0,1,2,0,1,0,1,1,ER
0,14,0,2,2,0,1,2,0,0,0,1,0,1,2,2,0,1,7,1,1,1,5,0,0,3,0,0,1,0,0,2,ER
,6,0,2,0,2,1,3,0,0,0,0,0,1,0,1,0,0,0,9,0,3,0,0,0,0,2,4,0,3,0,1,4,ER
,7,0,0,2,4,0,3,0,0,0,0,1,0,0,0,1,0,1,5,0,1,0,0,0,0,1,0,0,2,0,0,5,ER
,26,0,1,2,4,0,0,1,0,0,1,0,2,2,1,0,0,0,1,0,0,0,0,0,0,1,0,3,0,2,0,HM
,11,1,0,1,6,0,6,0,0,0,0,0,1,2,1,2,3,0,0,3,1,0,0,0,0,2,1,0,1,0,1,6,HM
,31,0,1,0,2,0,3,0,0,0,1,0,3,0,1,0,0,0,0,0,0,1,1,0,0,0,0,1,0,1,1,HM
,22,0,0,0,4,0,5,0,0,0,0,0,0,1,1,1,0,2,0,0,0,1,1,0,1,1,2,0,0,0,3,4,HM
,13,0,1,2,2,0,1,0,0,0,0,0,0,3,0,1,0,0,1,1,0,1,2,0,1,2,0,0,1,0,1,0,PC
,9,0,0,1,1,0,1,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,3,0,1,1,PC
,11,0,1,0,0,0,2,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,1,2,0,0,0,0,1,0,PC
,3,0,0,0,2,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,1,0,1,0,0,0,0,0,3,PC
```

Figura 3. Exemplo de um vetor de características

4. Resultados

Para realizar a validação da base de dados foi utilizado o ambiente WEKA[3] para a realização dos testes, realizando a classificação supervisionada em três tipos de classificadores distintos, que são: J48, Naive Bayes e Multilayer Perceptron.

No modelo apresentado, o processo de comparação é composto por duas fases, o treinamento e a classificação. Como a base de dados deste trabalho ainda é considerada pequena foi utilizada a técnica de *cross validation*[4].

Para cada classificador foram realizados 5 rodadas de testes, que incluíram todas as características, sendo que a base e as características utilizadas obtiveram dados promissores em identificação de autoria, entre 50-80%.

Ao final, foi desenvolvida a base de dados composta por 100 amostras de textos pertencentes a 25 autores consagrados da literatura brasileira. Essa base de texto foi coletada, tratada e validada em um período de 3 meses e será utilizada em testes em processos computacionais de identificação de autoria para fins de pesquisa no curso de Licenciatura em Informática da Universidade Tecnológica Federal do Paraná – Campus de Francisco Beltrão.

Foram gerados todos os vetores de características para validação da base, utilizando classes da gramática da língua portuguesa, tais como: advérbios, verbos, conjunções e pronomes. Tal base de dados será disponibilizada de forma pública em repositório da instituição para demais testes e desenvolvimento de aplicações que tenham cunho científico.

[3] Software de Mineração de Dados da University of Waikato - http://www.cs.waikato.ac.nz/ml/weka/

[4] Técnica estatística para estimar o desempenho de um modelo preditivo.

5. Conclusões

O objetivo principal deste artigo foi apresentar um método para coleta, tratamento e validação de uma base de textos. Foram coletadas 25 obras de 25 autores da literatura brasileira. A base de dados se mostrou robusta para a realização dos primeiros testes, porém ainda carece de estudos mais aprofundados para identificação das principais características dos textos. Em comparação com a literatura de identificação de autoria encontrada, o modelo apresentado apresenta uma contribuição para a área de reconhecimento de padrões, principalmente pela criação da base de dados que pode ser utilizada por outros pesquisadores, fazendo com que assim mais experimentos científicos envolvendo as características da língua portuguesa possam ser validados através da base de dados construída neste trabalho.

Como propostas de trabalhos futuros ainda serão aplicadas: o complemento da base de dados, chegando a 50 autores e 100 obras; e, o desenvolvimento de um repositório web para armazenamento e divulgação das pesquisas em base de dados textuais.

Referencias

Mcmenamin, G. R. (2002) "Forensic Linguistics - Advances in Forensic Stylistics". CRC Press, Florida-USA, 1a edition.

Pavelec, D. F. Justino, E. J. R. Oliveira, L. E. S. (2007) "Author Identication using Stylometric Features". Inteligencia Artificial, v. 11, p. 59-66.

Varela, P.J. Justino, E. J. R. Oliveira, L.E.S.(2011) "O uso de dicionário de atributos estilométricos na identificação de autoria de textos de língua portuguesa". The 8th Brazilian Symposium in Information and Human Language Technology. Cuiabá.

Varela, P.J. (2010) "O uso de atributos estilométricos na identificação de autoria de textos". Dissertação de Mestrado, PUC-PR, Curitiba.

Johnstone, B. (2005) Qualitative Methods in Socioliguistics. Oxford University Press, New York, p. 450.

Remco, R. Bouckaert, E.F. Mark A. Hall, G. H. Bernhard, P. Reutemann, P. Witten, I. H. (2010) "WEKA-experiences with a java open-source project". *Journal of Machine Learning Research*, 11:2533-2541.