Extended Abstract

Diz-me o que escreves, dir-te-ei quem és

Processamento de Língua Natural aplicado à literatura

Vanessa Alves Feliciano

Instituto Superior Técnico, Portugal

Abstract

The author identification tasks of a document have long been the target of the academic community interest.

The basis of this work is a framework developed by Nuno Homem, based on top-k most frequent words for each author. Our goal is to evaluate if the use of statistical data for each document and the top-k most frequent words, can improve the existing framework.

For the classification task of the documents it used the Weka. In addition, we evaluated the impact of excluding stop words from the list of most frequent words.

Finally, the application of the methodology was tested in the task of identifying other author attributes, such as: sex, birth century and decade of birth.

The results suggest that the use of statistical features, together with the top-k-used words, has improved the existing framework. Furthermore, it was observed that removing stop words of the most frequent words enhances the performance of this methodology. Finally, it was shown that it is possible to identify the sex of the author of a document and its century of birth. But when trying to identify the decade of birth the results are clearly below.

Keywords: Authorship attribution, features, Weka, stylometrics, stop words

1. Introduction

There are many documents with uncertain authorship. For example, “The Federalist Papers” [1] [2] [3] are a set of 77 political essays written by Alexander Hamilton, John Jay and James Madison, out of which some were claimed to be written by both Alexander Hamilton and James Madison.

Authorship Attribution can be interest in several areas. For example, in law I could be very useful to determine the authorship of an anonymous letter. [4]

This work is based on an existing framework created by Nuno Homem [5]. In this framework Nuno Homem uses a set of newspaper articles published in Público (a Portuguese daily newspaper).

Our main goal is to extend the existing framework and apply it to Portuguese books.

2. Related Work

2.1. Authorship Attribution using Stylometric Features

Using stylometrics for authorship attribution we assume that every writer has a particular (and unique) style which remains the same over time. [6]

Stylometrics feature are usually divided into 5 categories: lexical, structural, syntactic, content-specific and idiosyncratic [7] [8] [9].

Usually, the authorship attribution process has 3 steps [10]:

1. Gathering documents to be classified
2. Feature extraction
3. Document classification

2.2. Authorship Identification and Author Fuzzy "Fingerprints"

In this work [5] Nuno Homem try to identify an author using a fingerprint. A fingerprint is able to capture the characteristics of an author with a small probability of collision.

To create the fingerprint, Nuno Homem uses the top-k word frequencies in all identified texts.

When a new document arrives, a new fingerprint is calculated and the author with the most similar fingerprint is chosed.

2.3. Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging

In the work presented in [11], the author tries to extend traditional Authorship Attribution techniques, to identify people involved in chat conversations.

Each conversation was modeled as a sequence of "turns", i.e., a stream of symbols ad words typed consecutively by one subject.

Using the “turn” concept, conversational features were created: turn duration, writing speed, number of “return” characters and mimicry.

After calculate the individual performance for each feature the author achieved a list of the best features: # Exclamation Marks, # Emoticons, # Three points, # Uppercase letters, Turn duration, # Return chars, Words length, # Chars per second, # Question Marks, # Characters, Mimicry degree, # Words per second.

So, we can infer that having conversational features in count can improve traditional Authorship Attribution methods

2.4. Selecting Syntactic Attributes for Authorship Attribution

This work [12] presents a methodology to identify the most significant feature for documents written in Portuguese.

This work uses documents from 100 different authors divided in 10 subjects: Miscellaneous, Law, Economics, Sports, Gastronomy, Literature, Politics, Health, Technology and Tourism. Documents were gathered from 15 Brazilian newspapers.

For this study 4 types of features were used: conjunctions, adverbs, verbs and pronouns.

To evaluate the proposed methodology the author used an SVM classifier. Also, the author tried both an weighted-sum and a genetic algorithm. Using a weighted-sum approach, the best performance is achieved using 100 features. Using a genetic algorithm, the best performance is achieved using about 50 features.

Finally, only the set of features with the best performance was selected, and the error rate decreased from 42% to 26%. Table 1 contains the list of selected features.

Group	Quantity	Features
Adverbs	22	lá, dentro, adiante, em cima, ao lado, depois, sempre, com certeza, sem dúvida, ainda, quase, apenas, mais, todo, toda, bastante, nada, ninguém, nenhum, antes, qualquer, outro
Conjunctions	11	porém, por isso, assim como, que nem, segundo, embora, portanto, tais como, contanto que, de modo que, caso
Pronouns	10	seu, sua, quem, cujo, este, esta, o, a, aquele, onde
Verbs	15	ser, ver, pular, estar, ligar, estando, efetuando, fazendo, tendo, sendo, usando, pagando, aberto, visto, usado

1- Selected features

Analyzing the confusion matrix by subjects in % (Table 2) the recognition rate is about 86% and the worst subject is Miscellaneous, as expected.

	a. Miscellaneous	b. Law	c. Economics	d. Sports	e. Gastronomy	f. Literature	g. Politics	h. Health	i. Technology	J. Tourism
a.	82	7	1	2	1	2	1	2		1
b.	5	84	3		1	2	2	1		
c.	3	3	84			1	4	2		
d.	2		1	86	1	1	1	7	1	
e.		1		1	87	2	1	3		3
f.	4	3	2	1		87	3			
g.	1	1	6			3	88			
h.	1		2	4	3		1	88		2
i.	3	3	3			1	1		89	1
j.	1	1	2		6	1				89

2 - Confusion Matrix by subjects

2.5. Resources

2.5.1. Project Gutenberg

All documents used in this work were gathered from Project Gutemberg. This is an online library started in 1971 by Michael Hart.

This project provides full documents that are in the public domain.

2.5.2. Weka

Weka is a Data Mining tool developed by the Machine Learning Group from Waikato University, in New Zealand [13].

This tool, started in 1993, is a free software, under a GNU General Public License.

This software provides several Machine Learning algorithms that may be applied to a dataset or integrated in another software.

Weka also provides tool for: pre-processing, classify, clustering, association, feature selection and visualization.

Imported data are converted for an intermediate format named ARFF (Attribute Relation File Format).

An ARFF file is an ASCII file with 2 parts:

- Header: Define relation and attributes
- Data: List of instances, having the values for the defined attributes

Figure 1 represents an ARFF file.

```
@relation author.symbolic
@ATTRIBUTE NumberUtilization  NUMERIC
@ATTRIBUTE VowelUtilization  NUMERIC
@ATTRIBUTE class        {FranciscoJorgedeAbreu,JoséMartinianodeAlencar}
@DATA
0,37,FranciscoJorgedeAbreu
0,38,FranciscoJorgedeAbreu
0,37,JoséMartinianodeAlencar
0,38,JoséMartinianodeAlencar
0,37,JoséMartinianodeAlencar
```

1- ARFF file

3. Proposed Solution

3.1. Conceptual Model

Beside the top-k word frequencies, we decided to use a set of features to identify each document.

The new features are based on the statistical analysis of the document.

The features used in this work are: Vocabulary Richness, Sentences length, # Words per sentence, # Numbers, # Punctuation, # Vowels, # Characters (including white spaces), # Characters (excluding white spaces), # Periods, # Commas, # Question Marks, # Colons, # Semicolons, # Exclamation Marks, # Dashes, # Underscores, # Quotations, # Slashes.

Also, the top-10, top-25 and top-50 word frequencies were used.

To represent this data, in order to use Weka as a data mining tool, a structure was created (figure 2). For the top-k word frequencies the value is 1 if the word is part of the top-k word frequencies for all the documents and 0 otherwise.

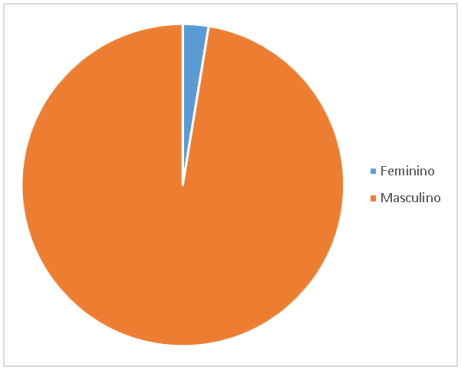
2 - Data structure

3.2. Implementation

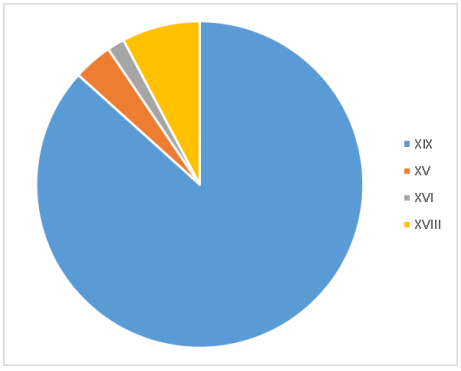
In this work, we used 233 books from 50 different Portuguese authors. The source for this documents was the Project Gutenberg.

The books' distribution by author is represented in Appendix A.

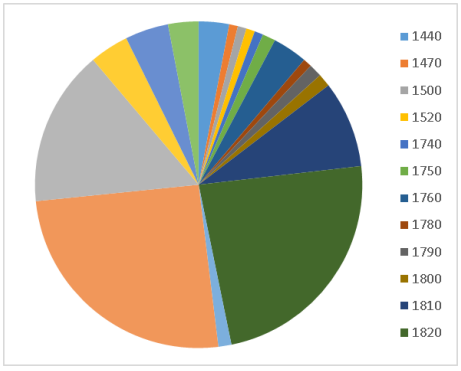
This documents are distributed by sex, century of birth of the author and decade of birth of the author as shown in the figures 3, 4 and 5, respectively.



3 - Distribution by sex of the author



4 - Distribution by century of birth of the author

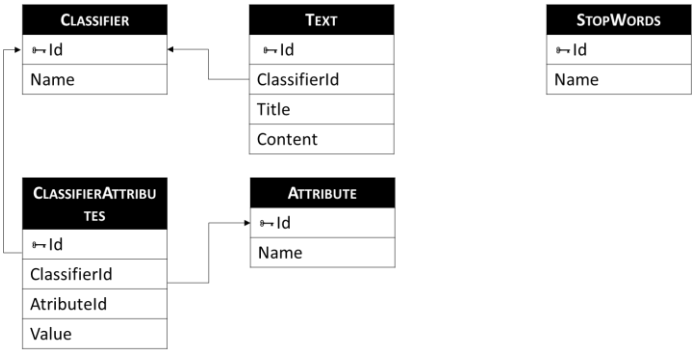


5 - Distribution by decade of birth of the author

In order to do an automatic analysis of this documents it was necessary to pre-process them. First, we removed the header and tail added by the Project Gutenberg. Then, we normalized the used encoding.

Finally, we observe that most of the documents were (wrongly) classified as “Camilo Castelo Branco”, because this was the author with more documents. So, we decided to extract 10 random text-parts from each author, in order to normalize the dataset.

To support this development, we created a SQL Database which architecture is represented in figure 6.



6 - Database architecture

Using this architecture we can classify the documents according to the author, but also according to several author attributes.

This database also contains a list of stop words[1] and several tables where intermediate information is saved during the process.

Finally, we developed several procedure in order to import documents into the database and extract feature data to create the ARFF file.

[1] https://code.google.com/p/stop-words

4. Evaluation

At the beginning, we tested the following scenarios: Only stylometrics features; Stylometric features + Top-10 word frequencies (including stop words); Stylometric features + Top-10 word frequencies (without stop words); Stylometric features + Top-25 word frequencies (including stop words); Stylometric features + Top-25 word frequencies (without stop words); Stylometric features + Top-50 word frequencies (including stop words); Stylometric features + Top-50 word frequencies (without stop words).

All this scenarios were tested with the Weka classifier, using this algorithms: Bayes Net, Complement Naïve Bayes, Naive Bayes, IB1, LWL, Bagging, Random Sub Space, Hyper Pipes, PART, FT, J48, Random Forest, Random Tree and Simple Cart. At the classification process we used cross-validation with 10 folds.

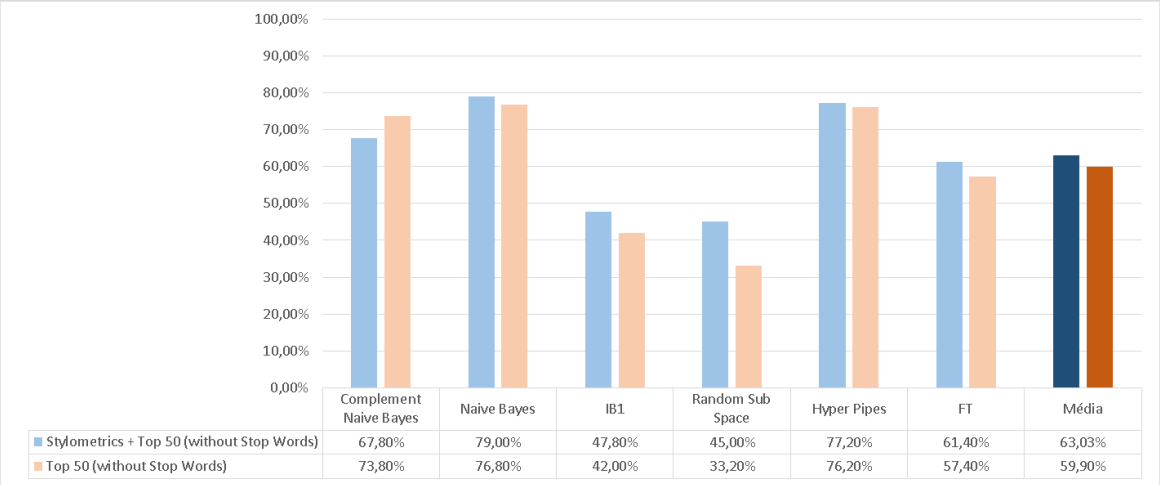
After this normalization, the algorithms that performed better were: Complement Naive Bayes, Naive Bayes, IB1, Random Sub Space, Hyper Pipes and FT.

Looking at the results (table 3), we conclude that using only the stylometric features we obtain the worst results.

	Average	Minimum	Maximum	Standard Deviation
Only stylometrics	32,77	4,00	47,60	9,77
Stylometrics + Top 10	33,47	13,80	51,20	8,57
Stylometrics + Top 25	39,96	21,20	64,20	9,92
Stylometrics + Top 50	45,54	19,40	71,00	12,25

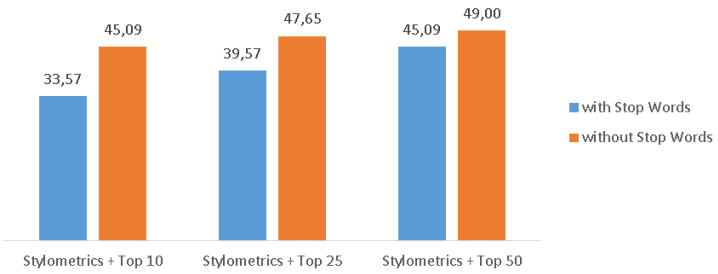
3 - Results by scenario

In order to evaluate if we improved the existing model by adding stylometric features, we decide to compare the version that performed better with and without this features. As we can observe ate figure 7, the results were improved in almost all algorithms and in average the accuracy rate increased 3,13%. So, we can conclude that using this stylometric featured we improved the exiting model.



7 - Stylometric features utilization

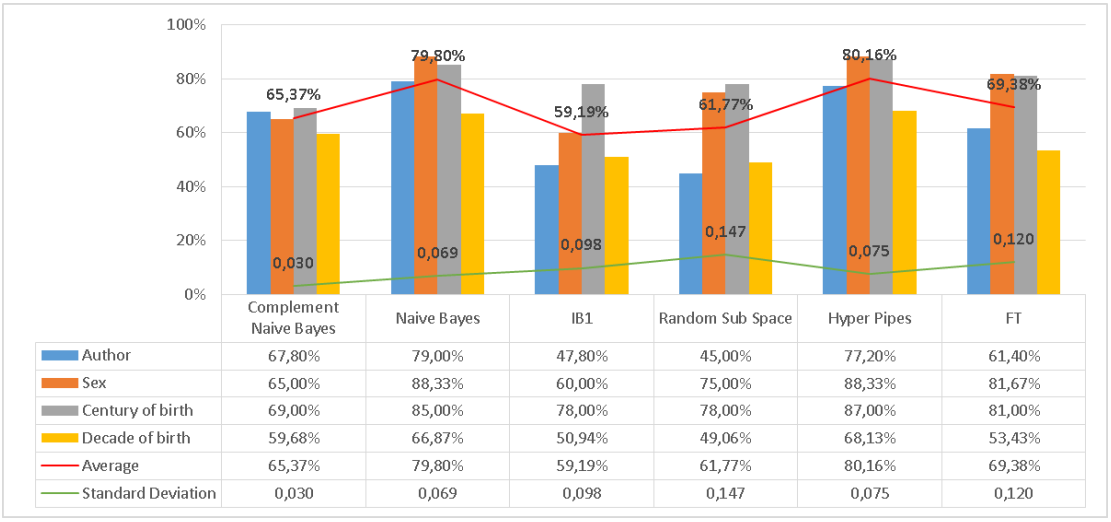
Then, we decide to evaluate if by removing stop words from the top-k word frequencies the results were improved. Using this technique the results were improved 7,7% in average (figure 8).



8 - Removing stop words

Finally, we decide to evaluate this methodology using the author attributes exiting in the database (sex, century of birth and decade of birth).

As we can see in figure 9, we can identify the sex and the century of birth with a performance similar to the previous results. But, when we try to identify the decade of birth of the author the results are considerably lower.



	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT
Author	67,80%	79,00%	47,80%	45,00%	77,20%	61,40%
Sex	65,00%	88,33%	60,00%	75,00%	88,33%	81,67%
Century of birth	69,00%	85,00%	78,00%	78,00%	87,00%	81,00%
Decade of birth	59,68%	66,87%	50,94%	49,06%	68,13%	53,43%
Average	65,37%	79,80%	59,19%	61,77%	80,16%	69,38%
Standard Deviation	0,030	0,069	0,098	0,147	0,075	0,120

9 - Identify author attributes

As a last experience, we also tried to evaluate if using only a subset of features will improve this model performance. So, evaluate the individual performance (see table 4) for each feature, we got the following scenarios:

- Using only features with performance above 6%
- Using only features with performance above 8%
- Using only features with performance above 2,5%

	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT	Average
Vocabulary Richness	2,00%	9,20%	6,80%	7,40%	3,60%	7,00%	6,00%
Sentences length	2,00%	10,80%	5,80%	13,00%	3,40%	11,80%	7,80%
# Words per sentence	2,00%	11,00%	8,40%	12,20%	3,60%	12,60%	8,30%
# Numbers	2,00%	4,00%	7,80%	7,40%	3,00%	3,80%	4,67%
# Punctuation	2,00%	12,00%	8,80%	13,00%	4,40%	12,80%	8,83%
# Vowels	2,00%	8,00%	2,80%	8,60%	2,20%	8,00%	5,27%
# Characters (including white spaces)	2,00%	6,00%	7,40%	5,60%	4,80%	6,80%	5,43%
# Characters (excluding white spaces)	2,00%	6,20%	6,40%	6,00%	4,80%	6,20%	5,27%
# Periods	2,00%	9,80%	6,60%	9,60%	3,00%	9,00%	6,67%
# Commas	2,00%	6,80%	5,80%	5,40%	5,20%	6,00%	5,20%
# Question Marks	2,00%	5,40%	7,60%	8,20%	2,60%	5,20%	5,17%
# Colons	2,00%	2,00%	2,00%	2,00%	2,00%	2,00%	2,00%
# Semicolons	2,00%	6,00%	10,40%	10,40%	3,00%	8,40%	6,70%
# Exclamation Marks	2,00%	10,60%	13,00%	14,40%	6,80%	11,80%	9,77%
# Dashes	2,00%	9,00%	6,60%	7,00%	5,60%	9,20%	6,57%
# Underscores	2,00%	4,60%	5,80%	5,80%	3,00%	6,60%	4,63%
# Quotations	2,00%	4,00%	4,00%	4,40%	3,80%	4,60%	3,80%
# Slashes	2,00%	2,60%	2,60%	2,00%	2,40%	2,00%	2,27%

4 - Features' individual performance

But none of these scenarios had better results than the initial results (using all features) as it is represented in table 5.

	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT	Average
Author	67,80%	79,00%	47,80%	45,00%	77,20%	61,40%	63,03%
Features with performance > 6%	72,20%	78,40%	43,00%	42,40%	78,40%	62,00%	62,73%
Features with performance > 8%	74,00%	78,20%	43,00%	36,60%	78,40%	57,60%	61,30%
Features with performance > 2,5%	67,80%	78,60%	47,80%	43,40%	77,20%	61,40%	62,70%

5 - Results with filtered features

So, it seems that although some feature have weak results they contribute to build a better model.

References

[1] R. M. Dabagh, "Authorship Attribution and Statistical Text Analysis," *Metodoloski zvezki,* pp. 149-163, 2007.

[2] S. R. Pillay, "Authorship Attribution of Web Forum Posts," *eCrime Researchers Summit (eCrime),* 2010.

[3] E. Stamatatos, "Automatic Authorship Attribution," *Conference of the European Chapter of the Association for Computacional Linguistics,* pp. 158-164, 1999.

[4] P. Juola, "A Prototype for Authorship Attribution Studies," *Lit Linguist Computing,* Junho 2006.

[5] N. Homem, "Authorship Identification and Author Fuzzy "fingerprints"," *Fuzzy Information Processing Society (NAFIPS), 2011 Annual Meeting of the North American,* 2011.

[6] Robert Layton, Paul Watters e Richard Dazeley, "Authorship Attribution for Twitter in 140 Characters or Less," *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second,* Julho 2010.

[7] O. Aslanturk, "Application of Cascading Rough Set-Based Classifiers on Authorship Attribution Application of Cascading Rough Set-Based Classifiers on Authorship Attribution," *2010 IEEE International Conference on Granular Computing,* 2010.

[8] N. Ali, "Evaluation of authorship attribution software on a Chat bot corpus," *2011 XXIII International Symposium on Information, Communication and Automation Technologies,* 2011 Outubro 2011.

[9] H. El-Fiqi, "A computational linguistic approach for the identification of translator stylometry using Arabic-English text," *2011 IEEE International Conference on Fuzzy Systems,* 2011.

[10] I. N. Bozkurt, "Authorship Attribution - Performance of various features and classification methods," *22nd IEEE International Symposium on Computer and Information Sciences,* pp. 1-5, 2007.

[11] M. Cristani, "Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging," *Proceedings of the 20th ACM international conference on Multimedia,* 2012.

[12] P. Varela, "Selecting syntactic attributes for authorship attribution," *The 2011 International Joint Conference on Neural Networks,* 2011.

[13] G. Holmes, "WEKA: A Machine Learning Workbench," *Proc Second Australia and New Zealand Conference on Intelligent Information Systems,* 1994.

Appendix A

Autor	Nº de Documentos	Nº médio de caracteres/documento	Nº médio de palavras/documento
Abílio Manuel Guerra Junqueiro	5	75897	12897
Adolfo Coelho	2	121559	19129
Alberto Leal Barradas Monteiro Braga	2	88739	15420
Alberto Pimentel	16	143932	24406
Alexandre Herculano	14	315536	55977
Ana de Castro Osório	2	198280	33804
Antero de Quental	10	48038	8067
António Augusto Teixeira de Vasconcelos	2	53343	9059
António Duarte Gomes Leal	4	69060	12176
Antonio Feliciano de Castilho	3	204893	35127
António Pereira Nobre	2	102231	18489
Augusto Gil	2	30225	5223
Camilo Castelo Branco	43	170240	29951
Carlos Testa	4	92509	15223
Eça de Queirós	9	582911	132204
Fernandes Costa	2	83438	13472
Florbela de Alma da Conceição Espanca	1	18874	3451
Francisco Jorge de Abreu	2	311606	50525
Gil Vicente	2	10264	1828
Gonçalo Anes Bandarra	2	37477	6535
Henrique Ernesto de Almeida Coutinho	2	9951	1678
Jaime de Magalhães Lima	11	143909	24634
João Augusto Marques Gomes	3	69393	11823
João Manuel Pereira Silva	2	206571	33432
João Marques de Carvalho	3	107898	17404
Joaquim Carlos Paiva de Andrada	2	86888	15018
José Agostinho de Macedo	3	60668	10379
José da Silva Mendes Leal	2	128381	21478
José Daniel Rodrigues da Costa	3	43614	7613
José Martiniano de Alencar	3	113704	19749
José Sobral de Almada Negreiros	6	25010	4486
Júlio Dinis	3	824104	141541
Luciano Cordeiro	3	82074	13120
Luís de Camões	2	416361	74512
Luiz Augusto Rebello da Silva	3	250542	68348
Manoel Caldas Cordeiro	2	37185	6156
Manuel Maria Barbosa du Bocage	5	19474	3321
Manuel Pinheiro Chagas	4	156502	27333
Maria Amália Vaz de Carvalho	3	303638	60453
Nicolau Tolentino de Almeida	2	84543	14651
Raimundo António de Bulhão Pato	2	64083	11280
Raul Germano Brandão	2	347551	88733

Rui de Pina	7	148940	32229
Sebastião de Magalhães Lima	8	117402	19076
Teixeira Bastos	3	66505	10605
Teixeira de Pascoais	4	19412	3452
Teófilo Braga	4	184989	31000
Vicente de Almeida de Eça	2	68148	11463
Visconde de João Batista da Silva Leitão de Almeida Garrett	3	186479	47617
Wenceslau José de Sousa de Morais	2	140996	23386

Books by author