Building a software model to preserve intellectual property and identify the authorship of a text

Sameer Saud Dakhel[1,*]

[1]*Computer Unit, College of Agriculture, Al Muthanna University.*

[*]*Corresponding Author:* samer_saoud@mu.edu.iq

Received 24March 2022, Accepted 7September2022, published 31December2022.

DOI: 10.52113/2/09.02.2022/11-15

Abstract : This article describes a software model algorithm that was implemented in order to determine the author of an unknown document based on the previously presented data and also studies the degree of accuracy of each of the measures for determining the authorship of a text. A comparative analysis is carried out based on the dependence of the quality of the identification of the text's creator on his stylistic affiliation.

The problem of determining the authorship of anonymous text in this programming model has been solved using a set of metrics, for each of which the following parameters are required: analysis of text frequency, total number of characters in the text, and dimension of the alphabet. These parameters are required for anonymous text and for the library of famous authors.

Keywords: authorship of a text, anonymous text, text frequency.

1. Introduction

To create an author library, all necessary parameters were calculated In advanced for each of them. In total, the library of this software contains 16 novel authors (Arabic literatures of the 19th and 20th centuries were selected, the library includes classic authors), as well as 8 texts of scientific style. Frequency analyses according to the scripts of these authors were previously calculated and stored in separate files in the project root folder.At the beginning of the program, the user must select the mode of operation: determining the author of the anonymous For the best work of this software, at the first stage, it is necessary to pre-process the text of the anonymous author [1]. All letters should be converted to the same case, all characters that do not belong to the alphabet, numbers and special characters should be excluded from the text, in addition, it is possible to remove spaces between words. The recommended options for clearing input text are: remove third-party characters, replace uppercase letters with lowercase letters, and have spaces between words. However, each user can choose the cleaning parameters at their discretion. The program also provides for adding a new author to the existing ones or updating the current library [2].

To determine the authorship of the text by the frequencies of the bigrams [3], first of all, it is necessary to know the number of characters in the input document. This value is determined at the stage of reading text from the file. You also need to know the power of the alphabet, that is, the number of letters in the alphabet. Initially, arrays are set to store the values of bigrams and the number of their repetitions in the text [4].

2. Methodology

2.1. Description of the working algorithm

Step 1. Before the start of the general cycle throughout the text, it is necessary to save the value of the first pair of letters of the text into an array of values, and set the number of repetitions of this bigram to 1.

Step 2. Organization of a general cycle throughout the text, starting with the second element

Step 3. If the next pair (the current and the following letter) is present in the corresponding array, then its number is increased by 1, and the algorithm goes to Step 2. If such a bigram is not present in the array, then it is stored in it, and its quantity in the required array becomes equal to 1, the algorithm goes to Step 2.

The algorithm ends its work when the cycle reaches the last pair of letters.

Sorting is carried out by the frequency of the bigrams, so that the most frequent bigram is in the first place in the array.

The frequency analysis of the text on the basis of grammes occurs in a similar way, only the work is carried out not with pairs of letters, but with each of them separately.

Next, each of the four measures is calculated. You will need values for the current known and anonymous author to calculate them. Bigram values and anonymous text frequency analysis are compiled once for each text, in accordance with the algorithms described above.

Next, a cycle is organized for all known authors and the calculation of the Khmelev measures, the Kullback divergence and the measure are carried out.

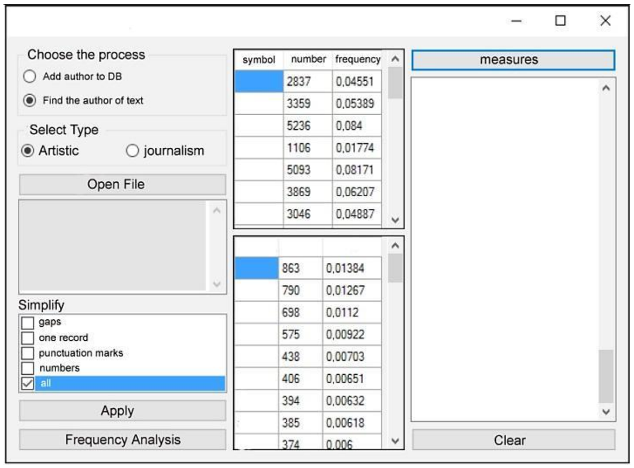
Step 1. The values of the frequency analysis and bigram values of the current famous author are read from the corresponding file and saved to the required array.

Step 2. There is a search for the bigram required by the algorithm, its frequency and the frequency response of the required single letter.

Step 3. Calculate the intermediate value of each of the measures and go to Step 2.

Step 4. After calculating the final value of the measures, the value of the current measure and the minimum value are compared. If the current value is less than the minimum, then it is necessary to reassign the values, and also keep the surname of the famous author by whom the comparison was made. Go to Step 1 until a comparison is made across the entire list of known authors from the program database.

Step 5. Displaying the text attribution results on the screen. An example is shown in Figure 1.



Choose the process
○ Add author to DB
● Find the author of text

Select Type
● Artistic ○ journalism

Simplify
□ gaps
□ one record
□ punctuation marks
□ numbers
☑ all

symbol	number	frequency
	2837	0,04551
	3359	0,05389
	5236	0,084
	1106	0,01774
	5093	0,08171
	3869	0,06207
	3046	0,04887

	863	0,01384
	790	0,01267
	698	0,0112
	575	0,00922
	438	0,00703
	406	0,00651
	394	0,00632
	385	0,00618
	374	0,006

Fig. (1): An example of how the attribution procedure works

3. Results and discussion

Further, this paper presents the results of experiments to identify the effectiveness of methods for determining authors of fiction and scientific literature of various sizes in the input file, as well as different volumes of well-known authors, whose texts are stored in the program database.

It should be noted that at the time of this writing, the program database contains 16 authors of fiction and 8 scientific literatures [5].

Table 1: Results of a comparative analysis (using the frequencies of the bigram).

authors	Anonymous text size					
	25 KB	50 KB	75 KB	100 KB	125 KB	150 KB
Khalid Al-Harbi	4406	3732	3381	3046	2759	2732
Daoud Sheikhani	4393	3879	3896	3614	3373	2940

The values of this measure of Khmelev using bigrams become more accurate with an increase in the size of the anonymous text. As a result of all experiments with this measure, the true author of the text was correctly identified[6].

Table 2: results of a comparative analysis according to the second measure of Khmelev (without the use of bigrams).

authors	Anonymous text size					
	25 KB	50 KB	75 KB	100 KB	125 KB	150 KB
Khalid Al-Harbi	124	109	70	8	67	65
Daoud Sheikhani	71	17	41	81	89	146

It should be noted that this measure is very unstable in its results and has a very low percentage of correctness of the author's definition (1 out of 12 experiments)[7].

Table 3: results of a comparative analysis by the values of the Kullback divergence.

authors	Anonymous text size					
	25 KB	50 KB	75 KB	100 KB	125 KB	150 KB
Khalid Al-Harbi	0,434	0,189	0,111	0,073	0,051	0,045
Daoud Sheikhani	0,394	0,171	0,116	0,079	0,062	0,043

Kullback divergence values become more accurate as the size of the anonymous text increases. As a result of all experiments with this measure, the author of the text was correctly identified [8].

Table 4: Results of a comparative analysis by the values of the measure X^2.

authors	Anonymous text size					
	25 KB	50 KB	75 KB	100 KB	125 KB	150 KB
Khalid Al-Harbi	32917	16458	10445	7607	5863	5035
Daoud Sheikhani	23556	11441	7982	5759	4706	3638

The values of the X^2 measure become more accurate as the size of the anonymous text increases. As a result of all experiments with this measure, except for a 25Kb literary text, the author of the text was correctly identified.

4. Conclusion

By applying the algorithm, a group of authors as well as a group of scientific texts were entered into the library of this program, where the frequency analysis according to the texts of these authors was previously calculated and stored in separate files in the root folder of the project. Good results have been obtained, and the authors have already been identified based on the data obtained as a result of the comparative analysis in this article, With the exception of the second Khmelev measure (which does not use bigrams), it can be concluded that all of the measures considered are highly effective on files of 25 KB or larger. And also in the meaning of the measures, there is no significant difference in which style is used to define.

References

[1] Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., and Steyvers, M., 2010, Learning author topic models from text corpora, ACM. Trans. Inf. Syst. 28 (1), 1-38.

[2] Prasad, R.S., U.V. Kulkarni, Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization, J. Comput. Sci. 6 (11), 1366-1376.

[3] Mansur, M., UzZaman, N., Khan, M., 2006, Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus, School of Engineering and Computer Science (SECS), BRAC University.

[4] Collins, M. J., 1996, A new statistical parser based on bigram lexical dependencies". Proceedings of the 34th annual meeting on Association for Computational Linguistics -. Association for Computational Linguistics, USA, 184-191. https://doi.org/10.3115/981863.981888

[5] Heafield, K., Pouzyrevsky, I., Clark, J. H., Koehn, P., 2013, Scalable modified Kneser-Ney language model estimation. ACL, 2, 690–696

[6] Witten, I. H., Bell, T. C., 1991, The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression, IEEE Trans. Inf. Theory 37 (4), 1085-1094.

[7] Sidorov, Grigori, 2013, Syntactic Dependency-Based n-grams in Rule Based Automatic English as Second Language Grammar Correction, Int. J. Comput. Linguistics Appl. 4, 169-188.

[8] Sanderson, C., Guenter, S., 2006, Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation. In Proceedings of the International Conference on Empirical Methods in Natural Language Processing, Comput Linguist Assoc Comput Linguist 482–491