Taras Shevchenko National University of Kyiv
Faculty of Computer Science and Cybernetics
Department of Computational Mathematics

GRADUATION THESIS
for a master's degree
in specialty 113 "Applied Mathematics"

on the topic:

Methods of author identification

of the 2 year student
Vladyslav Mykhailiuk

_____

Scientific advisor:
Doctor of Physical and Mathematical Sciences
Dmytro Klyushin

_____

The work was heard at a meeting of the Computational Mathematics and recommended for thesis defense, protocol № 9 of April 30, 2020.

Head of Department
of Computational Mathematics

Doctor of Physical and Mathematical sciences
S.I. Lyashko

Kyiv – 2020

Київський національний університет імені Тараса Шевченка
Факультет комп'ютерних наук та кібернетики
Кафедра обчислювальної математики

КВАЛІФІКАЦІЙНА РОБОТА
на здобуття ступеня магістра
за спеціальністю 113 «Прикладна математика»

на тему:

Методи ідентифікації авторів

студента 4 курсу
Михайлюка Владислава Юрійовича

_____

Науковий керівник:
доктор фізико-математичних наук
Клюшин Дмитро Анатолійович

_____

Робота заслухана на засіданні кафедри обчислювальної математики та рекомендована до захисту, протокол № 9 від 30 квітня 2020 р.

Завідувач кафедри обчислювальної математики — док. фіз.-мат. наук С.І. Ляшко

Київ – 2020

1 Contents

Contents

2 Introduction

Literature is often seen as a subject that concerns only reading and thinking. But, as is the case in almost all areas of intelligence, it intersects with others. In a work "Principles of Stylometry" in 1890, the Polish philosopher Vincent Lutoslawski used statistical approaches to construct a chronology of Plato's dialogues. Stylometry is said to have begun when, in 1851, Augustus de Morgan said about the texts of biblical authors that "one of the authors uses longer words." Stylometry is a linguistic discipline that applies statistical analysis to the literature by assessing the author's style using various quantitative criteria.

The development of computers and their ability to analyze large amounts of data creates opportunities for more sophisticated statistical methods. There are many works regarging the subject of literature texts analysis trying to solve the problem. Approaches such as Liuyu Zhou's in-depth study, NewsAuthorshipIdentificationwithDeepLearning, Mike Kestemont cite OverviewoftheAuthorIdentificationTaskatPAN2018, N. Smirnov cite Tableforestimatingthegoodnessoffitofempiricaldistributions, classical methods of machine learning Granichin cite WritingstyledeterminationusingtheKNNtextmodel, Grieve cite QuantitativeauthorshipattattAnevaluationoftechniques, and statistical methods Stamatatos cite Asurveyofmodernauthorshipattributionmethods, Giacomo Inches cite FindingParticiponshiphathat Shane Bergsma cite StylometricAnalysisofScientificArticles, Borisov LA cite IdentificationAuthorByFrequencies were considered in the works.

The issue of statistical analysis is in the interests of both literary critics and mathematicians. The tasks of clustering texts by author, genre, format, epoch of writing, or emotional coloring are interesting. Mathematical interest lies in the study of algorithms underlying the process of creative work of the brain. Also of independent value are the statistical methods of analysis of such multidimensional objects in their attributes, as literary texts written by professional writers.

The most popular task is to identify the authorship of the text. To solve it, we often use statistical methods based on the assumption that the writer in his works adheres to a certain behavior of writing, which forms certain invariants that emphasize the stylistic characteristics of the author's works. Such invariants can be quantitative features of the fraction of vowels and consonants in the text, the frequency of use of certain combinations of words, the frequency of marker words. Practical application of this question can be found in the search for plagiarism, linguistic, historical and criminological research.

This paper presents a comparative analysis of two approaches to iden-

tify the author. The first is based on comparing the frequencies of letter combinations. The second method hypothesizes that the text is written by a certain author and prove it using p-statistics cite AnAssumptioninMachineLearning. Also, since authors can write in different styles, the use of one standard for the author may not be effective. Therefore, the approach of clustering the author's texts and further finding a standard for each cluster is also considered. It is assumed that such an elalon corresponds to a particular style of the author. Clustering was performed by the hierarchical clustering method with the "author separation distance" $\hat{\rho}$, as a parameter.
In an alphabet with 26 characters, there is $26^3 = 17526$ different n -grams of length 3, which makes the calculation of the distance from author to author quite resource-intensive, so this paper proposes a method for finding the most effective n -grams to identify author.

The main purpose of this work is to compare methods in terms of resource utilization, computational speed, implementation complexity and identification accuracy. Also in this work the frequencies of n-grams are investigated as stylistic characteristics of the text.

3 Symbols and terms

A - number of authors
m - number of characters from the alphabet
The alphabet is further considered to be fixed.
n-gram - a sequence of n alphabet characters
K_α - number of texts of the author α
$N_{i,\alpha}$ - the number of letters in the i th text of the author α
n-PDF - the probability density function of n-grams of the text of the corresponding alphabet
$f_{i,\alpha}(j)$ - n-PDF of the i-th text of the author α, where the j corresponds to some n-gram, and varies from 1 to m^n
$f^k_{i,\alpha}(j)$ - n-PDF of the k-th partition of the i-th text of the author α, where the j corresponds to some n-gram, and varies from 1 to $a(n) = m^n$
$a(n) = m^n$ - quantity of n-grams of length n
$pstatistic(x, y)$ - measure of homogeneity of two samples

4 Methods

4.1 The density of the distribution function and its application to the author indentification

Let we have a library of texts of A authors. Let K_α be the quantity of texts of an author α, $N_{i,\alpha}$ - the number of letters in i-th text of author α. We assume that lengths of texts are large enough for the statistical analysis. We found for every text its presentation as n-gram frequencies and denote as $f_{i,\alpha}(j)$ corresponding n-PDF(probability density function) of the i-th text of the author α, and argument j coresponds to some n-gram and changes from 1 to $a(n) = m^n$: For each author, determine his weighted average PDF[2]:

$$F_\alpha(j) = \frac{1}{N_\alpha} \sum_{i=1}^{K_\alpha} f_{i,a}(j) N_{i,\alpha} \tag{1}$$

$$N_\alpha = \sum_{i=1}^{K_\alpha} N_{i,a} \tag{2}$$

These n-PDF will continue to play the role of author's etalons.
In 1, 2 the fact that the number of n-grams is less than the number of characters in the text is neglected, since $N_\alpha >> n$.
We introduce the biblical norm ρ_{ik}, as in the distance between n-PDF of

texts i and k:

$$\rho_{ik} = \|f_i - f_k\| = \sum_{j=1}^{a(n)} \mid f_i(j) - f_k(j) \mid \tag{3}$$

For every author, α build the probability density function $g_\alpha^+(\rho)$ of deviations $\rho_{i_\alpha,\alpha}$ of "his texts", and also the probability density function $g_\alpha^-(\rho)$ of deviations of texts of other authors to the author's etalon. Denote by $G_\alpha^\pm(\rho)$ corresponding cumulative distribution function. The minimal value of ρ for which $G_\alpha^-(\rho) = 1$ is denoted by ρ_α^+, and the maximum value of ρ for which $G_\alpha^($$\rho) = 0$ is denoted by ρ_α^-.
The meaning of the introduced notations is that all PDFs of texts of the author α are away by a distance of no more than ρ_α^+ from his average PDF F_α, and similarly all PDF of texts of other authors are away by not less than ρ_α^- from the avarage. The quantity $1 - G_\alpha^+(\rho_\alpha^-)$ is the propability to wrongly identify α as author of the text(type 2 error), and the quantity $G_\alpha^-(\rho_\alpha^+)$ is the probability to wrongly identify a text of an author α, as been written by other author(type 1 error). Denote by $G^+(\rho)$ - the cumulative distribution function of deviations of texts from the corresponding ethelons, and by $G^-(\rho)$ - the distribution of deviations of texts from "foreign"(other author's) ethelons:

$$G^+(\rho) = \frac{\sum_{\alpha=1}^{A} K_\alpha G_\alpha^+(\rho)}{\sum_{\alpha=1}^{A} K_\alpha} \tag{4}$$

$$G^-(\rho) = \frac{\sum_{\alpha=1}^{A} K_\alpha G_\alpha^-(\rho)}{\sum_{\alpha=1}^{A} K_\alpha} \tag{5}$$

Lets call the value of the division of authors the following value $\hat{\rho}$, for which the error of identification of the author of the text is minimal:

$$\hat{\rho_\alpha} = argmin(1 - G_\alpha^+(\rho) + G_\alpha^-(\rho)) = argmax(G_\alpha^+(\rho) - G_\alpha^-(\rho)) \tag{6}$$

$$\hat{\rho} = argmin(1 - G^+(\rho) + G^-(\rho)) = argmax(G^+(\rho) - G^-(\rho)) \tag{7}$$

The value $\hat{\rho}$ can serve as the upper level of text clustering.

Suppose we now have the text "0", the author of which must be identified. Then the author of the text is the author α, for which the norm $\rho_\alpha = \|f_0 - F_\alpha\|$ of difference between PDF $f_0(j)$ text "0" and the author's ethelon PDF $F_\alpha(j)$ is minimal:

$$\rho_\alpha = \|f_0 - F_\alpha\|, \quad \alpha^0 = \underset{\alpha}{argmin}\rho_\alpha^0 \tag{8}$$

The rule 8 is used only in case if $\min_\alpha \rho_\alpha^0 \leq \hat{\rho}$. If $\min_\alpha \rho_\alpha^0 > \hat{\rho}$ it will be decided that there are no possible authors in the library.

4.2 Method using p-statistics

4.2.1 P-statistics

P-statistics - the probability of a given statistical model for which, provided that the null hypothesis is true, the statistical sums will be the same or have more extreme values than for the actual results.
Suppose we have two samples $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_m)$ from the general populations X and Y respectively. For the sample z, it is known that it belongs to X or Y - the task is to identify to which set z belongs. We hypothesize H about the homogeneity of two samples from general populations with distribution functions $F_x(u)$ and $F_y(u)$, respectively, $x = (x_1, x_2, ..., x_n) \in X$ and $y = (y_1, y_2, ..., y_m) \in Y$ - test samples in which $x_{(1)} \leq ... \leq x_{(m)}$ and $y_{(1)} \leq ... \leq y_{(m)}$. Let the hypothesis H be $F_x(u) = F_y(u)$, then according to $A(n)$ Hill's assumption:

$$p(y_{(k)} \in (x_{(i)}, x_{(j)})) = \frac{j-i}{n+1}, i < j \tag{9}$$

Using a sample of $y = (y_1, y_2, ..., y_m)$, we can calculate the frequency h_{ij} of the random event $y_{(k)} \in (x_{(i)}, x_{(j)}))$ and calculate the confidence interval I_{ij} for the probability $p(y_{(k)} \in (x_{(i)}, x_{(j)}))$ with a given level of significance level β. Denote by L the number of intervals for which $\frac{j-i}{n+1} \in I_{ij}$ holds. Then, we define the degree of homogeneity of the samples x and y as the proportion of intervals for which $\frac{j-i}{n+1} \in I_{ij}$ among all intervals is true:

$$h_{xy} = \frac{2L}{n * (n-1)} \tag{10}$$

Since, h_{xy} is the random event frequency $\frac{j-i}{n+1} \in I_{ij}$ with probability $1-\beta$, we can construct a confidence interval I_{xy} for events $\frac{j-i}{n+1} \in I_{ij}$ with significance level β. If $1-\beta \in I$ then the hypothesis H is confirmed, otherwise rejected. The value h_{xy} is a measure of the homogeneity of the samples x and y. By swapping x and y and finding the frequency h_{yx} and the confidence interval I_{yx}, we can construct another test to test the hypothesis H. Since the measure h_{xy} is not symmetric, we can construct a symmetric measure of homogeneity:

$$h = pstatistic(x, y) = \frac{1}{2}(h_{xy} + h_{yx}) \tag{11}$$

4.2.2 Adaptation of P-statistics to identify the author of the text.

Suppose we have a library from the previous method with K_α works by the author α, α varies from 1 to A. We will divide each text in the library into

K parts, and for each part we will find $f^k_{i,\alpha}(j)$ - n-PDF of the k-th part of the text. Denote by $g_{i,\alpha(j)}(j)$ the set of frequencies of the j-th n -gram of the j-th text of the author α:

$$g_{i,\alpha}(j) = \{f^1_{i,\alpha}(j), f^2_{i,\alpha}(j), ..., f^K_{i,\alpha}(j)\} \tag{12}$$

Then we introduce the distance between the texts a and b, as:

$$\|f^{(\cdot)}_a - f^{(\cdot)}_b\| = 1 - \frac{\sum_{j=1}^{a(n)} pstatistic(g_a(j), g_b(j)}{a(n)} \tag{13}$$

5 Computational experiments

5.1 Development tools

The text library was downloaded from https://www.gutenberg.org/, which contains more than 60,000 books. The program for downloading the library is written in the programming language NodeJS v.12.13.1.
The algorithm for tests is written in programming language C++ 17. Parallel streams were used for increase of speed of calculations.
jupyter python was used to visualize the data.
Testing was performed on a computer with the following specifications:

- OS: 64bit Windows 10 Pro
- Central processor: Processor Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz, 2801 Mhz, 4 Core(s), 8 Logical Processor(s)
- RAM: 32.0 GB

5.2 Data-in

Testing was performed on a set of texts by 16 authors: George Manville Fenn, Sir Walter Scott, R.M. Ballantyne, U.S. Copyright Office, Robert Louis Stevenson, Jules Verne, W.H.G. Kingston, George Sand, Anthony Trollope, Charles Dickens, G. A. Henty, Mór Jókai, Fergus Hume, Alexandre Dumas, E. Phillips Oppenheim, William Le Queux. Each author has at least 50 books of 200,000 characters in length in the library. The tests were performed using 5,000, 10,000, 20,000, 50,000, 100,000, 200,000 first characters of the texts. The texts of each author were divided into a training and test sample of 25 texts in each.
Since the calculation of distances for trigrams is resource-intensive, the most "significant" trigrams were selected, ie those that most distinguish one author from another. The following statistics were chosen as a measure of "influence":

$$v(j) = \frac{\sum_{\alpha=1}^{A} D(f_{\cdot,\alpha}(j))}{D(f_{\cdot,\cdot}(j))} \tag{14}$$

$$D(f_{\cdot,\cdot}(j)) = \frac{\sum_{\alpha=1}^{A} \sum_{i=1}^{K_\alpha} (f_{i,\alpha}(j) - M(f_{\cdot,\cdot}(j)))^2}{\sum_{\alpha=1}^{A} K_\alpha} \tag{15}$$

$$M(f_{\cdot,\cdot}(j)) = \frac{\sum_{\alpha=1}^{A} \sum_{i=1}^{K_\alpha} f_{i,\alpha}(j)}{\sum_{\alpha=1}^{A} K_\alpha} \tag{16}$$

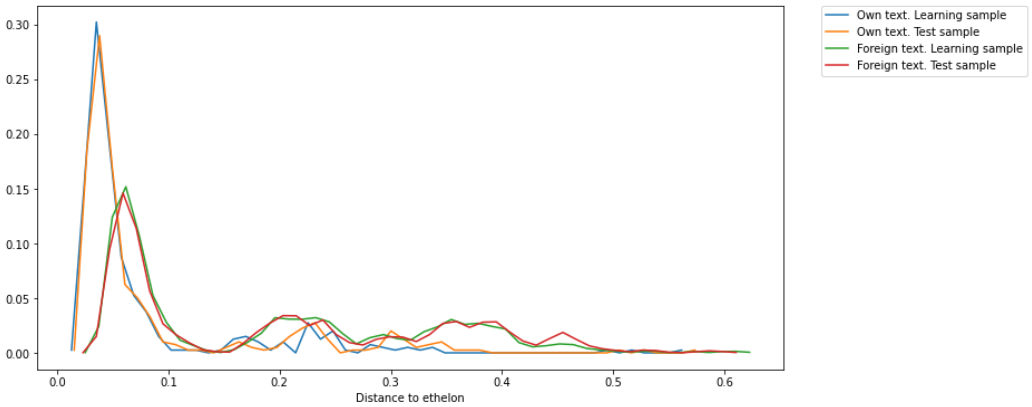

Figure 1: Density of distance distribution. Monograms.

$$D(f_{\cdot\alpha}(j)) = \frac{\sum_{i=1}^{K_\alpha}(f_{i,\alpha}(j) - M(f_{\cdot\cdot}(j)))^2}{K_\alpha} \tag{17}$$

$$M(f_{\cdot\alpha}(j)) = \frac{\sum_{i=1}^{K_\alpha} f_{i,\alpha}(j)}{K_\alpha} \tag{18}$$

$D(\cdot)$ - the variation, $M(\cdot)$ - the expectation. We get that $v(j)$ - shows what part of the general variation of frequencies j-th n-gram is the variation of these frequencies local to the authors. If $v(j)$ is close to 0 - this means that when you change the author of the text, this frequency changes much more than when you change the text to the text of the same author. Among all the 17526 trigrams, 1802 was chosen with a degree of "significance" less than 0.6.

5.3 Method using the density of the distribution function

Tables 1, 2, 3 contain information about the average distances of texts to the standards of each author for monograms, bigrams, trigrams, respectively. And the figures 1, 2, 3 show the densities of the distances to the ethelons. From the obtained tables we see that Jules Verne, Charles Dickens and Alexandre Dumas - have a fairly large average distances to their own ethelons, this is due to the fact that these authors write in quite different styles.

Author	The average distance of own texts to the ethelon		The average distance of other people's texts to the ethelon	
	Training sample	Test sample	Training sample	Test sample
George Manville Fenn	0.0338465	0.0340254	0.148186	0.160377
Sir Walter Scott	0.0284826	0.0663677	0.134049	0.147205
R.M. Ballantyne	0.0256525	0.0259007	0.132092	0.146816
U.S. Copyright Office	0.0474249	0.0434929	0.230926	0.241641
Robert Louis Stevenson	0.0660438	0.0500744	0.13504	0.150064
Jules Verne	0.194515	0.232865	0.214297	0.223655
W.H.G. Kingston	0.0310675	0.0344322	0.140399	0.153805
George Sand	0.0345426	0.0521552	0.345482	0.354845
Anthony Trollope	0.0312901	0.0393127	0.144474	0.157774
Charles Dickens	0.137723	0.201273	0.147524	0.156707
G. A. Henty	0.0299935	0.0315441	0.141307	0.155087
Mór Jókai	0.23157	0.221696	0.266655	0.265208
Fergus Hume	0.0346095	0.0355302	0.132482	0.146382
Alexandre Dumas	0.10102	0.152488	0.283127	0.291188
E. Phillips Oppenheim	0.0273646	0.0319321	0.136483	0.150508
William Le Queux	0.0312424	0.0289607	0.131785	0.145958

Table 1: Comparison of distances to ethalons. Monograms.

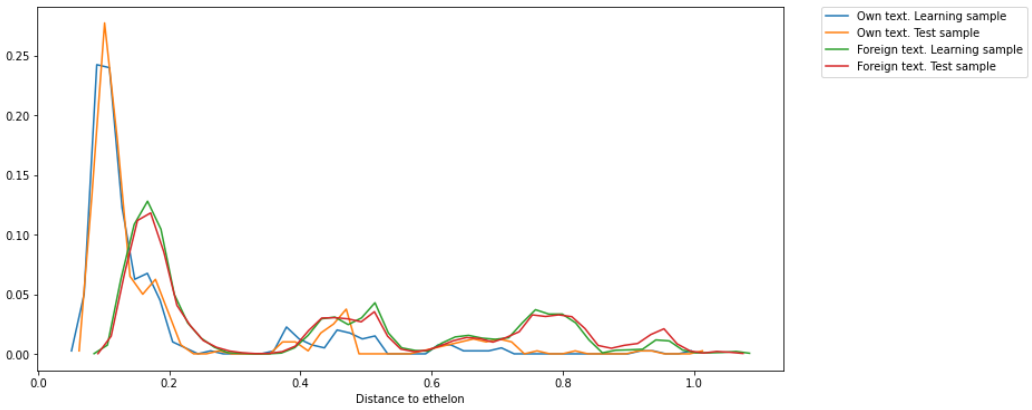

Figure 2: Density of distance distribution. Digrsams.

Author	The average distance of own texts to the ethelon		The average distance of other people's texts to the ethelon	
	Training sample	Test sample	Training sample	Test sample
George Manville Fenn	0.0994828	0.100713	0.337656	0.364495
Sir Walter Scott	0.0857056	0.160734	0.312519	0.337815
R.M. Ballantyne	0.0842319	0.0831463	0.303563	0.333081
U.S. Copyright Office	0.142319	0.131085	0.518372	0.536536
Robert Louis Stevenson	0.153465	0.126039	0.305945	0.336367
Jules Verne	0.434099	0.507919	0.467984	0.481571
W.H.G. Kingston	0.0986361	0.101119	0.319829	0.347821
George Sand	0.0953532	0.131873	0.720437	0.736496
Anthony Trollope	0.0978328	0.120733	0.335552	0.363841
Charles Dickens	0.289434	0.424478	0.322943	0.339882
G. A. Henty	0.0958608	0.0976739	0.32082	0.349905
Mór Jókai	0.473379	0.450451	0.564311	0.563759
Fergus Hume	0.103849	0.110262	0.308154	0.336897
Alexandre Dumas	0.228394	0.34131	0.612664	0.625097
E. Phillips Oppenheim	0.0855664	0.0948044	0.316374	0.344707
William Le Queux	0.0926607	0.0948494	0.306627	0.334875

Table 2: Comparison of distances to ethelons. Digrsams.

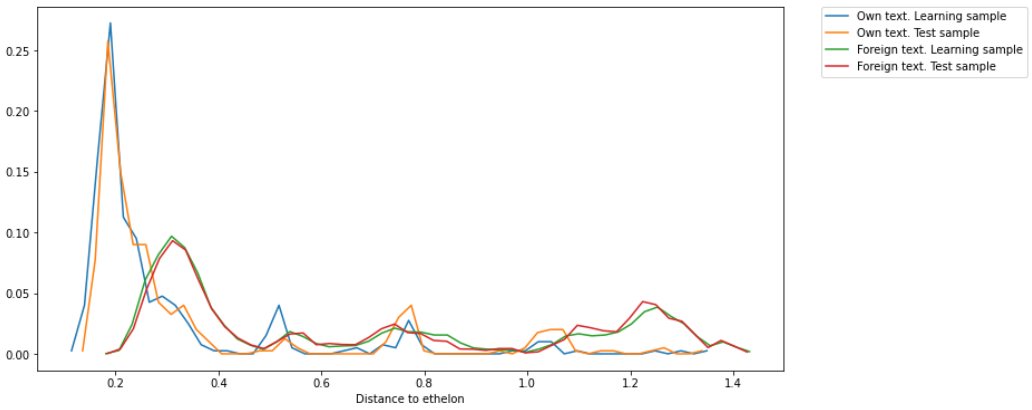

Figure 3: Density of distance distribution. Trigrams.

Author	The average distance of own texts to the ethelon		The average distance of other people's texts to the ethelon	
	Training sample	Test sample	Training sample	Test sample
George Manville Fenn	0.180418	0.18861	0.596302	0.623662
Sir Walter Scott	0.171463	0.283575	0.524401	0.549622
R.M. Ballantyne	0.172195	0.169814	0.51445	0.545968
U.S. Copyright Office	0.283822	0.276685	0.815953	0.833536
Robert Louis Stevenson	0.259684	0.228195	0.511306	0.545252
Jules Verne	0.655446	0.770669	0.789989	0.801207
W.H.G. Kingston	0.194357	0.204415	0.539513	0.568905
George Sand	0.162869	0.224804	1.15696	1.16934
Anthony Trollope	0.185929	0.224242	0.572796	0.603712
Charles Dickens	0.466389	0.668956	0.530902	0.546114
G. A. Henty	0.189261	0.19251	0.539592	0.571259
Mór Jókai	0.632626	0.749113	0.671995	0.671159
Fergus Hume	0.192996	0.206572	0.530133	0.561459
Alexandre Dumas	0.368591	0.561702	1.02425	1.02995
E. Phillips Oppenheim	0.170172	0.19042	0.546378	0.576854
William Le Queux	0.194836	0.198912	0.520049	0.550307

Table 3: Comparison of distances to ethelons. Trigrams.

	Training sample	Test sample
monograms	0.695	0.6875
digrams	0.7375	0.765
trigrams	0.7725	0.785

Table 4: Accuracy for text of 200,000 characters. Method using the density of the distribution function

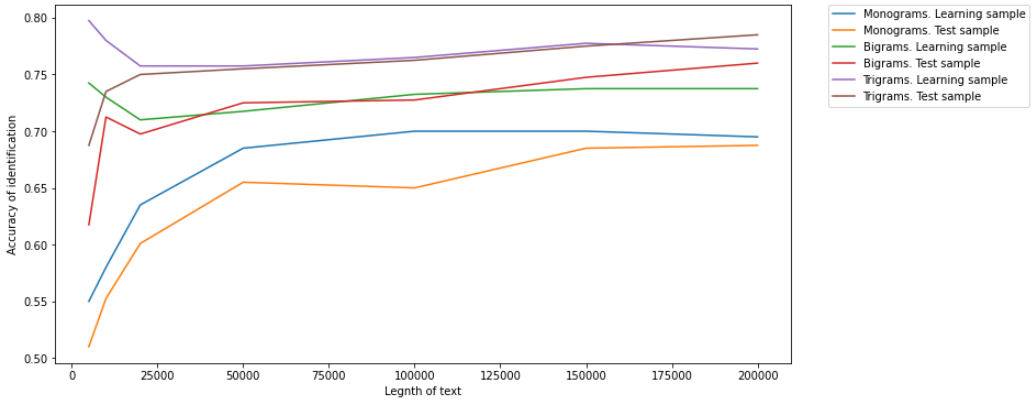

Figure 4: Identification accuracy for different lengths of texts.
Figure 5 shows the accuracy of identifying authors for different lengths of n-grams.

Figure 4 shows the accuracy of identifying authors for different lengths of texts. You can see that for the accuracy of identification of texts using bigram and trigrams on the test sample in some places exceeds the accuracy of identification on the training sample. This is due to the fact that the test sample included narrower texts in their style.

The following conclusions can be drawn: - Starting with a text length of 20,000 characters, the change in the accuracy of identification accuracy for bigram and triram is insignificant (1-2%), and starting with 50,000 characters, the change in error for monograms is also insignificant.

5.4 Method using p-statistics

Figure
reffig:6 shows the dependence of the accuracy of the author's identification

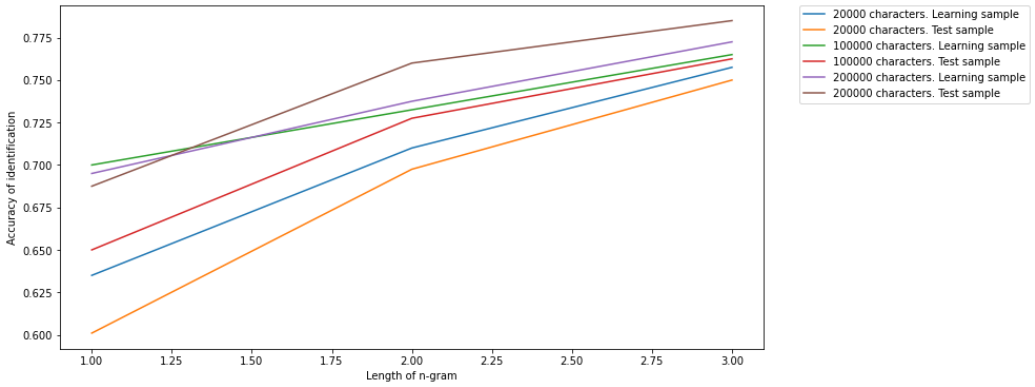

Figure 5: Identification accuracy for different n-gram lengths

	Training Sample	Test Sample
monograms	0.7725	0.7075
digrams	0.83	0.8055
trigrams	0.8325	0.815

Table 5: Accuracy for text of 200,000 characters. P-statistics.

depending on the length of the text for K = 20. We observe that in general, with increasing length of texts, the accuracy of identification increases. There is also a gap of 5-10 % accuracy between test and training samples.

5.5 Clustering

Since authors can write in different styles, clustering author's texts with subsequent finding of an ethelons for each author can be quite effective. The hierarchical clustering method was chosen as the clustering method, as the maximum distance between clusters can be transferred to the method. The authors' separation measure was used as a parameter (Tables 6, 8, 10 for monograms, bigrams and trigrams, respectively). Tables 7,9, 11 contain the number of clusters by authors for monograms, bigrams and trigrams, respectively.

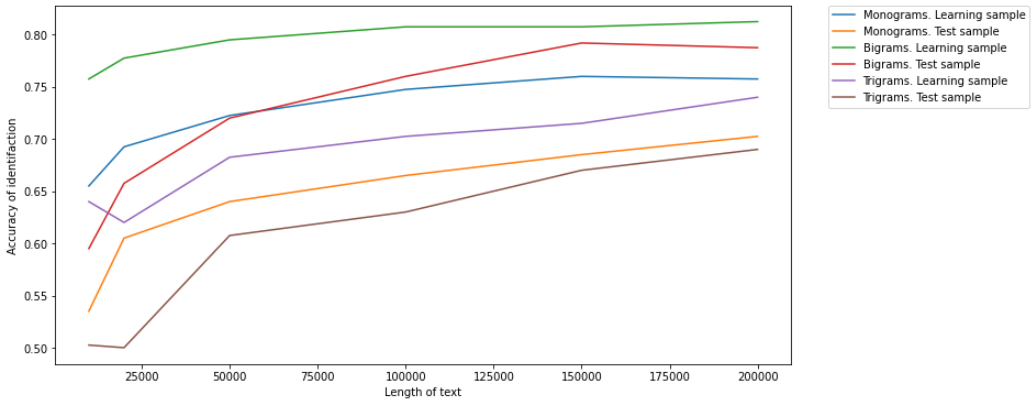

Figure 6: Identification accuracy for different lengths of texts. K = 20

Figure 6 shows the dependence of the accuracy of identification depending on the number of parts into which the text was divided. As the number of pieces of text increases, the accuracy of identification by the accuracy of identification by trigrams decreases sharply, which means that the length of the text is insufficient to obtain high accuracy. As the number of pieces of text increases, the identification accuracy for bigram and monogram decreases slightly. For Bigram the most effective is the division K = 7, for monograms K = 15, for trigrams K = 3. Figure 8 shows the dependence of the identification accuracy for the corresponding partitions. We see that for trigrams, even when divided into 3 parts of 200,000 characters in the text is not enough to reach around the statistical limit.

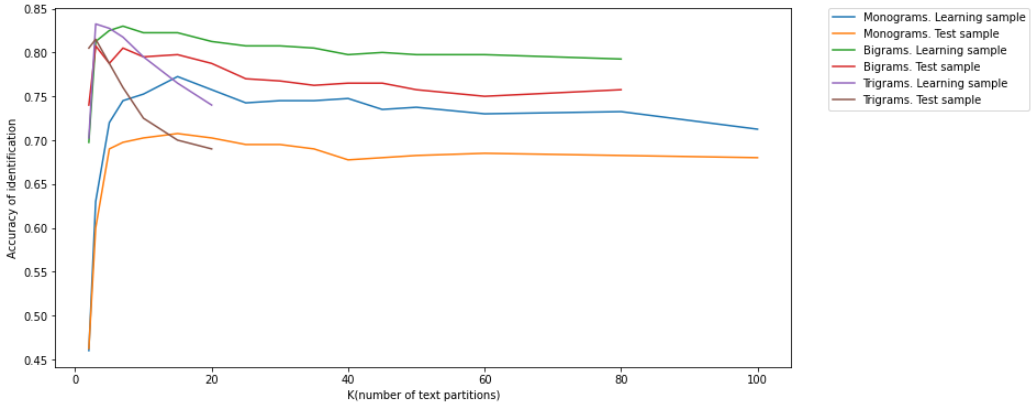

Figure 7: Accuracy of identification of various partitions. Text length 200000

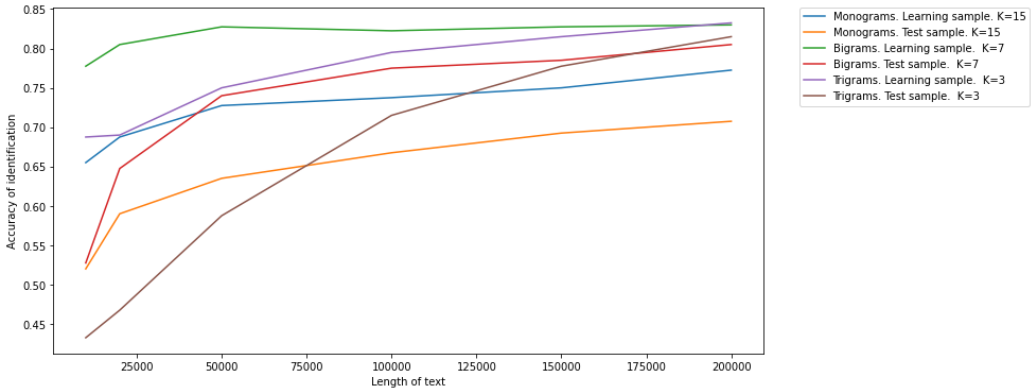

Figure 8: Identification accuracy for different lengths of texts. K = 15 for monograms, K = 7 for bigrams, K = 3 for trigrams

Author	$\hat{\rho_\alpha}$ for method using distribution density	$\hat{\rho_\alpha}$ for method using p-statistics
George Manville Fenn	0.0463045	0.187912
Sir Walter Scott	0.0375501	0.158608
R.M. Ballantyne	0.0318236	0.143223
U.S. Copyright Office	0.0657424	0.176374
Robert Louis Stevenson	0.0543053	0.209341
Jules Verne	0.178568	0.365568
W.H.G. Kingston	0.0397744	0.157326
George Sand	0.0549845	0.249267
Anthony Trollope	0.0426615	0.172711
Charles Dickens	0.0823775	0.224908
G. A. Henty	0.04078	0.177473
Mór Jókai	0.247522	0.335897
Fergus Hume	0.0374213	0.170879
Alexandre Dumas	0.0929762	0.259524
E. Phillips Oppenheim	0.0411207	0.17619
William Le Queux	0.0387453	0.172344

Table 6: $\hat{\rho_\alpha}$. Monograms

Author	number of clusters for method using density distribution function	number of clusters for method using p-statistics
George Manville Fenn	6	4
Sir Walter Scott	4	6
R.M. Ballantyne	5	3
U.S. Copyright Office	1	1
Robert Louis Stevenson	5	8
Jules Verne	4	4
W.H.G. Kingston	4	7
George Sand	1	1
Anthony Trollope	5	3
Charles Dickens	3	5
G. A. Henty	2	1
Mór Jókai	2	2
Fergus Hume	5	9
Alexandre Dumas	2	2
E. Phillips Oppenheim	1	1
William Le Queux	5	6

Table 7: Number of clusters. Monograms

Author	$\hat{\rho_\alpha}$ for method using distribution density	$\hat{\rho_\alpha}$ for method using p-statistics
George Manville Fenn	0.0463045	0.187912
Sir Walter Scott	0.0375501	0.158608
R.M. Ballantyne	0.0318236	0.143223
U.S. Copyright Office	0.0657424	0.176374
Robert Louis Stevenson	0.0543053	0.209341
Jules Verne	0.178568	0.365568
W.H.G. Kingston	0.0397744	0.157326
George Sand	0.0549845	0.249267
Anthony Trollope	0.0426615	0.172711
Charles Dickens	0.0823775	0.224908
G. A. Henty	0.04078	0.177473
Mór Jókai	0.247522	0.335897
Fergus Hume	0.0374213	0.170879
Alexandre Dumas	0.0929762	0.259524
E. Phillips Oppenheim	0.0411207	0.17619
William Le Queux	0.0387453	0.172344

Table 8: $\hat{\rho_\alpha}$. Bigrams

6 Analysis of results

6.1 Execution time

The method using the density of the distribution function turned out to be much faster than the method using p-statistics, due to the fact that for p-statistics you need to calculate $\frac{K(K-1)}{2}$ confidence intervals for each n-gram, compared with the sum of the frequency differences n -gram.

Author	number of clusters for method using density distribution function	number of clusters for method using p-statistics
George Manville Fenn	1	1
Sir Walter Scott	2	3
R.M. Ballantyne	3	3
U.S. Copyright Office	1	10
Robert Louis Stevenson	6	7
Jules Verne	4	2
W.H.G. Kingston	6	4
George Sand	1	1
Anthony Trollope	1	3
Charles Dickens	3	4
G. A. Henty	1	1
Mór Jókai	2	1
Fergus Hume	13	1
Alexandre Dumas	2	2
E. Phillips Oppenheim	1	1
William Le Queux	7	7

Table 9: Number of clusters. Bigrams

Author	$\hat{\rho_\alpha}$ for method using distribution density	$\hat{\rho_\alpha}$ for method using p-statistics
George Manville Fenn	0.236771	0.269608
Sir Walter Scott	0.24953	0.266926
R.M. Ballantyne	0.200519	0.258232
U.S. Copyright Office	0.363961	0.275157
Robert Louis Stevenson	0.237309	0.264058
Jules Verne	0.724427	0.349427
W.H.G. Kingston	0.209656	0.268868
George Sand	0.221383	0.338605
Anthony Trollope	0.246643	0.27525
Charles Dickens	0.312767	0.285331
G. A. Henty	0.225755	0.274972
Mór Jókai	0.508067	0.321032
Fergus Hume	0.231127	0.264058
Alexandre Dumas	0.333833	0.348594
E. Phillips Oppenheim	0.218364	0.270533
William Le Queux	0.22903	0.26859

Table 10: $\hat{\rho_\alpha}$. Trigrams

6.2 Identification accuracy

For monograms, bigrams and trigrams, the method using p-statistics gives better results by $3-4\%$ (for texts longer than 50,000). However, for small texts, p-statistics give worse results than the method using the density of the distribution function.

Using clustering, the accuracy in the test sample increased by approximately 5% for monograms, and by approximately 10% for bigrams and trigrams in the case of the density distribution method, and the best accuracy of 91.75% was obtained for trigrams. In the case of the p-statistic method, the accuracy for monograms and trigrams increased by approximately 5%, and for trigrams almost did not change, and the best accuracy of 85.25% was obtained for bigrams.
From the obtained results it is possible to draw conclusions: - in the case of clustering, the method using the density of the distribution function gives better results. But the results of two methods lye very close to each other. Therefore, it can be said that probality distribution of some n-gram does not contain more information about author style that the mean of distribution. - In this context, based on the previous point, P-statistics can be considered as

Author	number of clusters for method using density distribution function	number of clusters for method using p-statistics
George Manville Fenn	1	1
Sir Walter Scott	2	21
R.M. Ballantyne	3	4
U.S. Copyright Office	1	9
Robert Louis Stevenson	6	7
Jules Verne	3	4
W.H.G. Kingston	6	5
George Sand	1	3
Anthony Trollope	1	1
Charles Dickens	3	5
G. A. Henty	6	1
Mór Jókai	2	7
Fergus Hume	2	12
Alexandre Dumas	2	2
E. Phillips Oppenheim	1	1
William Le Queux	8	5

Table 11: Number of clusters. Trigrams

	Training Sample	Test Sample
monograms	0.9125	0.75
bigrams	0.98	0.8925
trigrams	0.99	0.9175

Table 12: Accuracy of the method using the density of the distribution function with clustering

	Training Sample	Test Sample
monograms	0.97	0.8125
bigrams	0.91	0.8525
trigrams	0.9875	0.805

Table 13: Method using p-statistics. K = 15 for monograms, K = 7 for bigrams, K = 3 for trigrams

some empirical version of the analogy of the confidence interval, as without clustering p-statistics gave better results, and with clustering - worse. - the distribution of n -grams really changes with the change of style, however, for greater accuracy, you need to look for other markers

7 Conclusions

As a result of the research, 2 methods of identification of an unknown author of a work belonging to the library of known authors were implemented, the method of text clustering was also implemented and testing of methods with and without clustering was performed. A criterion method was also proposed to select the n-grams that would best serve as a marker to identify the author. 800 texts by 16 authors were used for testing. As a result, it was found that the method that uses the density of the distribution function is suitable for identifying the authors of works of both large texts (50,000+ characters) and small (10,000+ characters). And the method that uses p-statistics is only suitable for use on large works. With clustering of texts, much better results were obtained in a test sample for both methods.

8 References

[1] Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. 2017.

[2] Marcia Fissette. Author identification in short texts. 2010.

[3] J. Grieve. Quantitative authorship attribution: An evaluation of techniques. 2007.

[4] Giacomo Inches, Morgan Harvey, and Fabio Crestani. Finding participants in a chat: Authorship attribution for conversational documents. 2013.

[5] Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, and Walter Daelemans. Overview of the author identification task at pan-2018. 2018.

[6] D.A. Klyushin, S.I. Lyashko, and S.S. Zub. Author identification in short text.

[7] Borisov L.A., Orlov Yu.N., and Osminin K.P. Identification of the author of the text by the distribution of frequencies of letter

combinations/Идентификація автора текста за розподілом частот буквосполучень. 2013.

[8] Granichin O., Kizhaeva N., Shalymov D., and Volkovich Z. Writing style determination using the knn text model. 2015.

[9] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics, 19, 1948.

[10] Stamatatos. A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 2009.

[11] Liuyu Zhou. News authorship identification with deep learning. 2016.