Scientific Visualization, 2025, volume 17, number 2, pages 110 - 122, DOI: 10.26583/sv.17.2.08

Analysis of the Error Structure in Identifying the Author of a Text Using the Nearest Neighbor Graphs

M.Yu. Kislitsyna[1]

Keldysh Institute of Applied Mathematics RAS

[1] ORCID: 0000-0002-2542-8914, voronina.miu@yandex.ru

Abstract

In this paper, the nearest neighbor graph method is used to analyze the relationship between a large number of multidimensional vectors, which represents the distribution of letter combinations (n-grams) in the text, where n is 3. The task is the authorship attribution problem, which belongs to the field of natural language processing. The graph of the nearest neighbors is built according to the pattern distribution of the authors and visualizes the points of concentration and sparsity, which allows to identify the structure of text classification errors. The corpus consists of more 8 thousand authors and more 100 thousand literary texts in Russian including translations. Thus, this is one of the most extensive experiments with literary texts in Russian. All authors have at least five works in the corpus, each of which contains more than 10 thousand letters. The author's pattern is calculated by averaging the 3-gram statistics of his texts. The error structure associated with the proximity of texts and authors to the average pattern of lexicon is determined using graphs. It is shown that the densest centers of the graph are close to the average lexicon pattern with varying degrees of proximity. The text recognition error of such authors is about two times higher than the error of authors who are far from the lexicon. Some literary genres, such as philosophical ones, are localized at special distances.

Keywords: The authorship attribution problem, nearest neighbors graph, an author.

1. Introduction

In the paper the author recognition problem of literary text in Russian is considered. This task belongs to the natural languages processing (NLP) problem and is a rapidly developing section of the application of artificial intelligence technologies. There are several examples include monographs [1-3], scientific publications [4-10], and also proceedings of specialized conferences [11-15]. Today, amount of works in NLP and text recognition are increasing and the number exceeds thousands. However, most of the works have a result which can't be scaled to other similar sets of texts or corpus. The main reason is using incomplete corpus that doesn't include the general distribution of target classes. Usually, a typical corpus contains several dozen authors and from several hundred to ten thousand texts. Because of limited sampling it is not possible to fully categories the types of recognition errors.

In cluster analysis method of the nearest neighbor graphs is commonly used [16-20]. Nearest neighbor graph method is heuristic algorithm of clustering data in some metric space. However, it may not be effective for an incomplete corpus, as it requires a large number of data points. In this paper, nearest neighbor method is used for visualizing the connections between a large number centers of groups. It allows to detect the concentration and sparsity of groups and analyze a data configuration. As a result, this will help to determine the behavior of the classification method in relation to the data.

In the previous work [21], an attempt was made to collect as much complete corpus of literary texts in Russian as possible. Article considers the result of authorship attribution in the corpus using the nearest neighbor method. The text features are 3-gram letter combination and an author pattern, which is calculated from averaging of 3-gram distributions of authors texts. The classification error is determined from the micro-average accuracy and is calculated using the entire cross-validation approach. In this article the graph vertex is the author patterns which used in [21]. The study demonstrates that the configuration of authors within the corpus is not random. Furthermore, visual analysis revealed well-separated node structures that remain stable when varying the neighborhood degree from 1 to 20. The patterns within these structures occupy specific distance regions relative to the pattern of the entire corpus. Additionally, the texts by the authors within these structures have different recognition properties.

2. Statistical analysis of the corpus

The corpus of texts in Russian contains 108,518 texts and 8,287 writers. Authors and texts in the corpus are chosen by requirements: each text must be at least 10 thousand characters in length, each author must have more 5 texts, genres such as non-fiction, poetry, plays and religion are not included in the corpus. The requirement of text length is followed from estimating the sufficient sample size for the text feature vector. Filtering by genre is done to avoid using text with special words and terminology, as well as those with a specific rhythm, such as poetry. In [21], the filtration procedure is described in more detail.

The corpus structure by origin languages of texts is 66,215 texts (61% of total number) in Russian, and other 42,303 texts (39%) in translated from other languages. Proportion of Russian-speaking and other language speaking authors in the corpus is same ratio as the texts. The total volume of the analyzed works is approximately 34 billion characters, of which 20 billion correspond to texts in the original Russian language. Thus, Russian-language and translated works, both in total length, and in terms of the number of texts correspond approximately as 60:40.

The text feature vector is frequency 3-gram character. Before the feature extraction, each text preprocess by removal all symbols (and gaps) except characters in Russian. After that 3-gram characters frequency is calculated for such “solid” text. Further statistical characteristics of texts in the corpus are set for preprocessed “solid” texts. Let $D_a^i(j)$ is the empirical frequency of 3-ram of character j in i-th text of author a and N_a^i is total number of characters in i-th work. Let n_a is a number of author’s works. Then the author pattern $F_a(j)$ is determine by average 3-gram frequency of own texts by formula:

$$F_a(j) = \frac{1}{N_a}\sum_{i=1}^{n_a} N_a^i D_a^i(j), \quad N_a = \sum_{i=1}^{n_a} N_a^i \tag{1}$$

The distance between texts and authors is considered as distance between the corresponding distributions in the L1 norm. The authorship attribution algorithm was tested using the cross-validation method, where distance between an author and own text is calculated with removing the text feature from the author’s pattern.

Finally, the distance from a text to pattern of the corpus is given by the formula

$$z_{ab}^i = \frac{1}{1-\delta_{ab}N_b^i / N_b}\sum_{j=1}^{J}\left|D_a^i(j) - F_b(j)\right| \tag{2}$$

where δ_{ab} is the Kronecker symbol, $J = 32^3 = 32\,768$. Formally, for error-free classification it is required that

$$\forall i,a,b:\ z_{aa}^i < z_{ab}^i,\ b \neq a \tag{3}$$

The error is calculated from micro-average accuracy. In terms previous formula, the error is the ratio of the number of violations of inequality (3) to the total number of texts. According to the results of the numerical experiment in [21], the error of classification by the nearest neighbor method with 3-gram feature vector in L1 norm is 23.8%. At the same time, Russian-speaking and other languages speaking authors are classified with the error 19.6% and 30.1% correspondently. A number of correctly recognized authors (with zero error) is 2,191 (about 26% of total number of writers) and consist of both Russian-speaking and translated writers. On the other hand, 472 authors (about 5%) are entire unclassified.

The statistical characteristics of number and average length of texts by authors in the corpus are presented in Fig. 1. The left graph on Fig.1 shows the distribution of authors by the number of their texts. It can be seen, that major part of the authors has few texts. There are more than 1400 authors (about 17% of total) have five texts and 4986 authors (about 60% of total) have less ten texts. Dependency between number of authors and their texts average length is shown on the right graph of Fig. 1. The mode of the distribution is about 300 thousand letters.

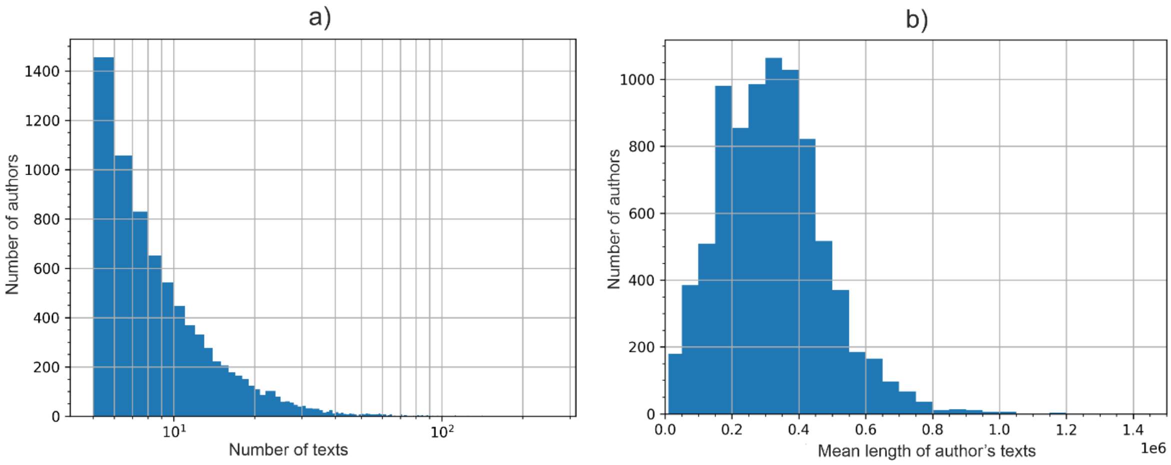

Fig. 1. Distribution of the authors by the number of texts (a) and by average length of author's texts (b).

An author with the maximum number of texts per author is Barbara Cartland with 374 texts. She usually writes in romance novels genre. Note, that only 27 of 374 (about 7%) authors were identified incorrectly, despite the fact that they were translated by different translators. Next top ten authors in the corpus with the largest number texts are: Sergey Zverev (327 texts, 160 errors), Daria Dontsova (266, 2 errors), Robert Stein (256, 44 errors), Friedrich Neznansky (241, 61 errors), Vladimir Kolychev (225, 4 errors), Daria Kalinina (217, 1 error), Alexander Tamonikov (214, 99 errors), Elena Arsenyeva (205, 83 errors), Chingiz Abdullayev (183, 9 errors). Part of the authors are well classified with error about 7% of their own text number. Other of them have the error of more than 30%. Therefore, a large number of texts is not provided well recognition.

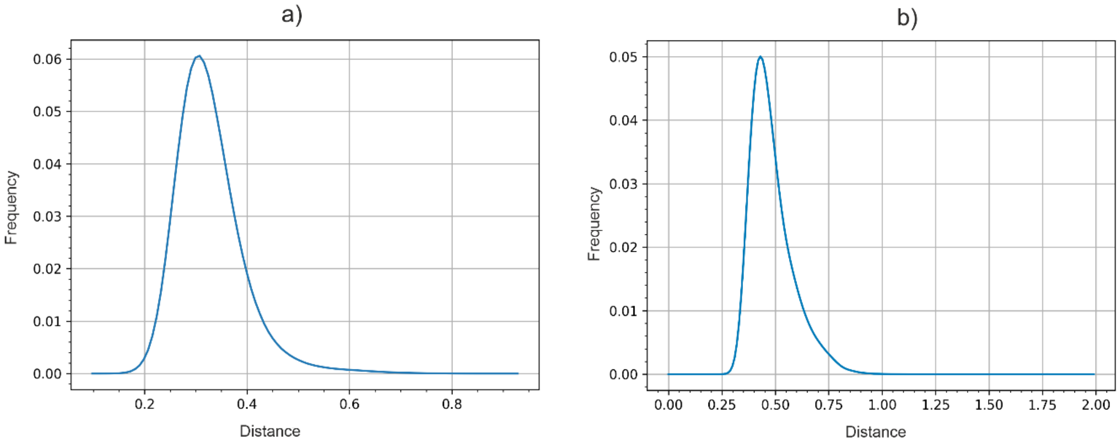

Fig. 2. Distance distribution between text-text (a) and author-author (b).

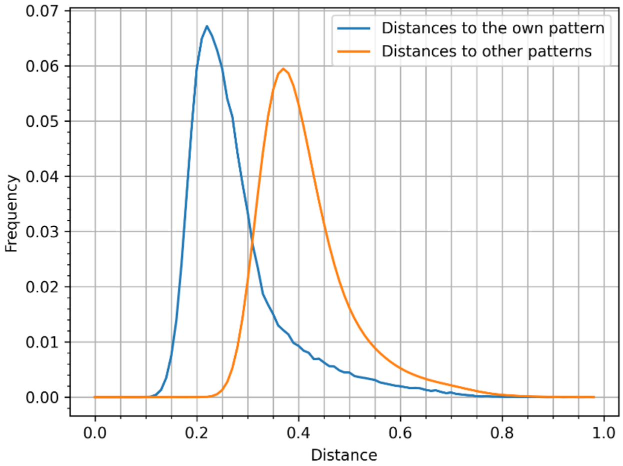

Fig. 3. Distance distribution from text to own pattern (blue line) and other patterns (red line).

Fig. 2 shows that half width at half maximum of the distances distribution between texts is about twice large than patterns. Hence, using data of author's text separately for the nearest neighbor method is likely less effective in comparison with classification by author's pattern. In practice, the classification error by text-to-text distances is larger than text-to-pattern distances. For this corpus the text-to-text recognition error is about 0.3. It can be seen from Fig. 3 that distances from texts to own authors generally are less than similar distances to other authors, which provides a low value of the error. The most common distances between the patterns and between the texts are about 0.3 and 0.4 correspondingly. The mode of distribution of text-to-own author distances is 0.22 and text-to-other author is 0.37.

Due to the nearest neighbor method doesn't require pre-training and is realized by exhaustive search the classification algorithm doesn't include any random initial parameter. Therefore, the error of classification depends on configuration of patterns in the corpus. If the configuration is random, then the error cannot be corrected within this approach. In [22, 23] a criterion of data dependency was formulated by using the random nearest neighbor graph to a set of distances. In the next section, the configuration of patterns is analyzed according to the criterion and the k-nearest neighbor graph (for k from 1 to 20) is built.

3. The structure of the nearest neighbor graphs for author's patterns

In this section some statistic properties of the pattern configuration are discovered. The method of detecting dependences between the patterns was proposed in [22, 23]. In this method is used the directed nearest neighbor graph. An algorithm of building the graph and calculation his statistics includes definition an adjacency matrix. The adjacency matrix is based on the matrix of distances between the patterns. For the nearest neighbor graph, the positions of the smallest distances in each row of the matrix correspondent nonzero elements in the adjacency matrix. The edge is oriented from *i*-th row to the *j*-th column with nonzero value in the adjacency matrix. A similar algorithm is performed for the k-nearest neighbor graph (k-nn graph) by selecting the next after the smallest distances in the rows. All calculation is realized in Python 3.8 programming language. The visualization of the graphs is made by library NetworkX [24] and it is constructed up to isomorphism. Figures 4-7 below show the nearest neighbor graphs of orders 1, 2, 3 and 20.

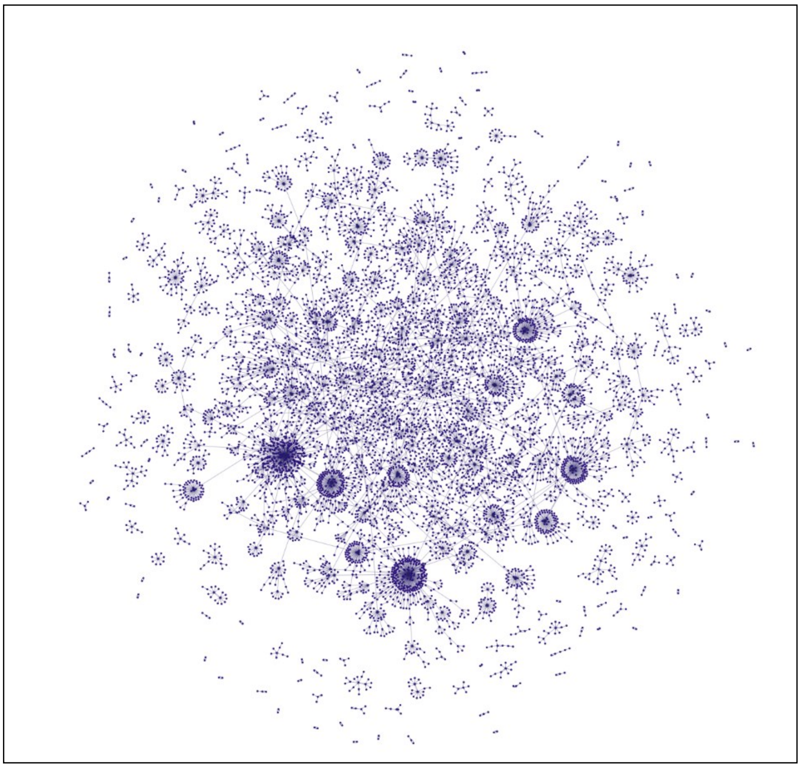

Fig. 4. The nearest neighbor graphs of orders k=1

In Fig. 4 is shown the nearest neighbor graph. The graph consists of 284 fragments. If the patterns were independent random vectors, then the most likely number of fragments would be about 2000 with the range about 1000. The lower amount of disconnected fragments illustrates the some connectivity forces between the elements. Despite of the normality of the connected distribution components for random data, the probability of data independence (for considered system) leads to zero.

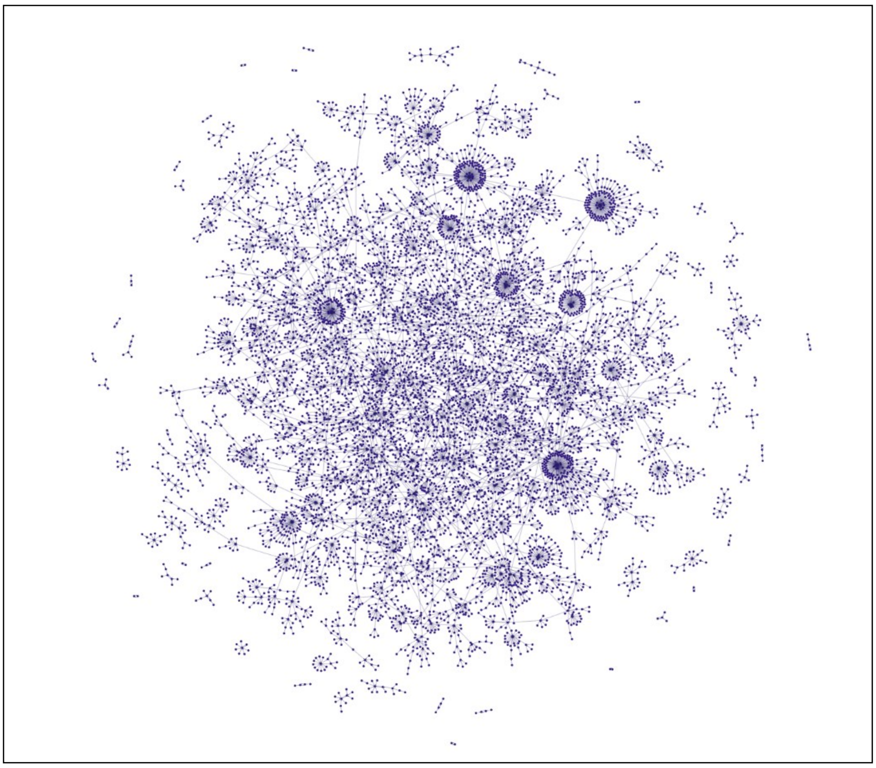

Fig. 5. The nearest neighbor graphs of orders k=2

The graph of the second-order neighbors consists of 117 fragments (Fig. 5). The connectivity is enhanced by including more longer cycles. However, the maximum degree values are lower than for the case k=1.

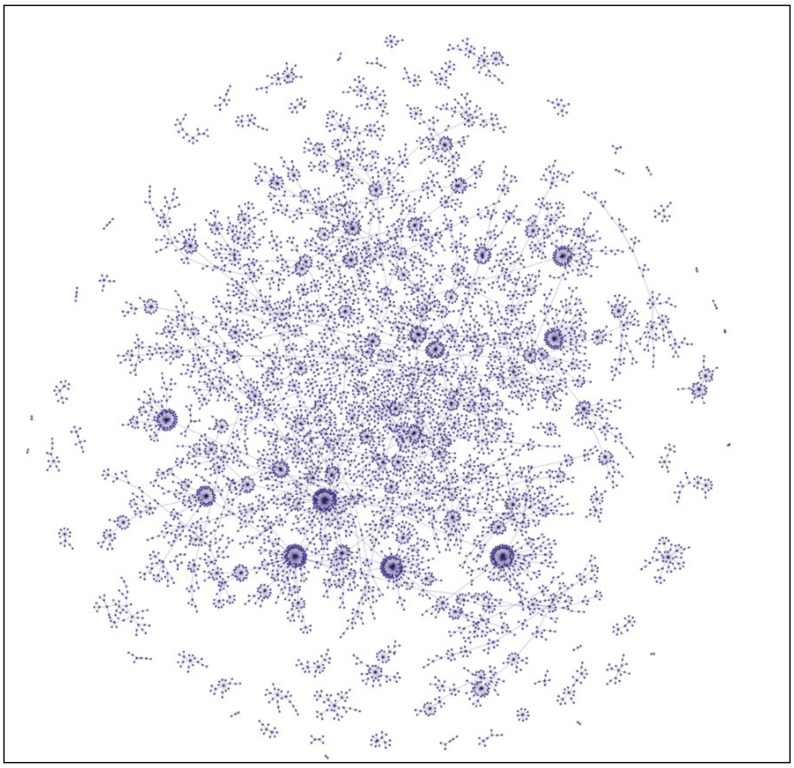

Fig. 6. The nearest neighbor graphs of orders

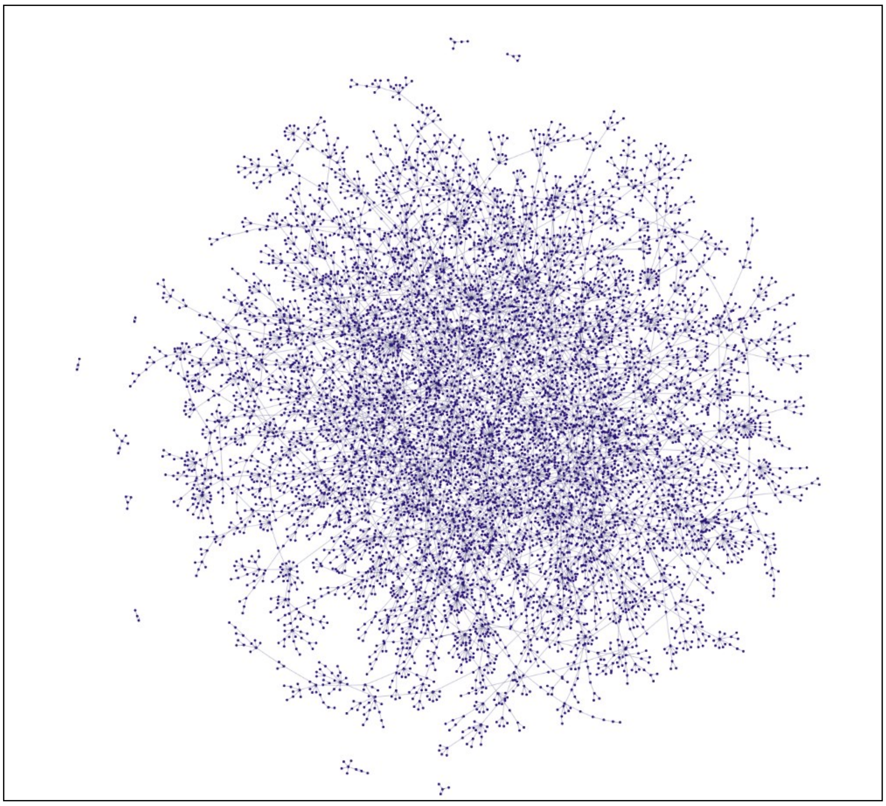

Fig. 7. The nearest neighbor graphs of orders 20

Fig. 6 and Fig. 7 show the third-order and twenty-order neighbors' graphs, respectively. When k=3, the graph includes 93 fragments and is even more connected than the second-order graph. The increase in connectivity follows a monotonic trend up to the sixth-order neighbor, after which the value of fragment amount fluctuates chaotically while maintaining a general tendency towards decrease (see Fig. 8 below). The twenty-order graph includes 23 fragments.

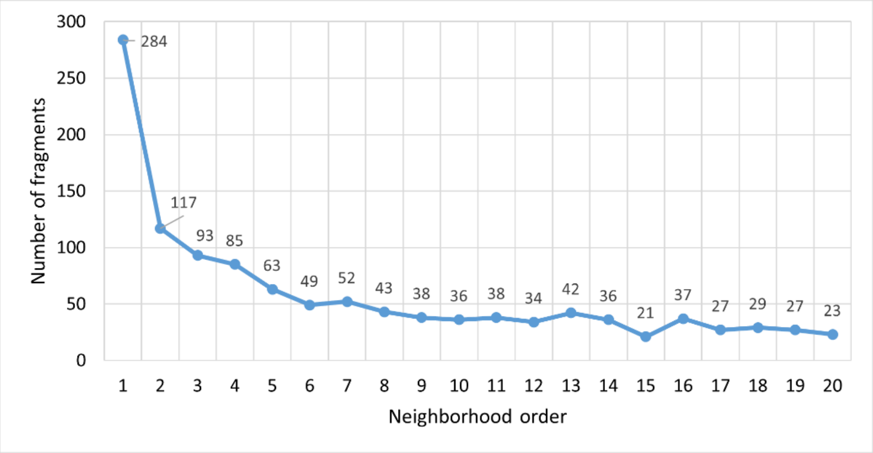

Fig. 8. The nearest neighbor graphs of orders 20

The structure of the k-neighbor graphs with k from 1 to 20 correspond to a system of vectors that is dependent with a probability close to one. According to [23], the structure is independent with probability less than 10^{-6}. Therefore, the nodes with large degree value are not random and can have an objective reason relate to patterns configuration. Note that in addition to reducing the number of fragments, the number of maximum degree values has also decreased.

Let's select the first hundred authors with the maximum degrees for values of k from 1 to 20. It turns out that the sets of authors for the different values of k have non-empty intersection and equal 33. Further, that 33 patterns are used for creating correction algorithm of the classification. On the other hand, there are 2,249 patterns which haven't incoming edges (leaf) in the whole k-near cases.

From the graphs structure it follows that patterns with maximum degree correspond to authors without strong individual differences and are close to the 3-gram distribution of the general lexicon. Patterns with zero degree in any k-near neighbors likely consists of texts with special thematic or individual writing manners. Consequently, corpus authors who are close to the average value are likely to have a greater error in classifying their texts than the texts of authors with a zero degree. This idea is confirmed below. It should be emphasized that this hypothesis follows from a visual analysis of the graphs of k-nearest neighbors. In this situation, visualization of the calculation result is important for understanding the properties of the array, especially for large datasets, when sorting by their numerical parameters is not particularly informative.

4. The dependence of the author's recognition error on the distance to the corpus

Let's analyze the text classification error of the authors from founded in previous section graph structure - the 33 authors with stable high degree and 2, 249 patterns without incoming edges. It was noted, that the 33 authors have summary error of their texts equals 38% (total the corpus error 23.8%). That more than twice of the classification error of the entire corpus. In 25% cases of wrong recognition the nearest to text pattern is one of the 33 authors and in 20% cases is one of a pattern, whose have edges with the 33. Moreover, such authors with edges more likely misidentification with one of them (36% cases). The authors without incoming edges also have the error, which more than the total error, and equals 30%. But only

0.7% of them misrecognized with other authors without connections and about 10% with the neighbor to them patterns from 1 to 20. Hence, wrong identification texts of such far from other patterns authors doesn't depend on pattern neighborhood. In opposite, the concentration points and their first neighbor are more likely misrecognition with each other and have high error value of their text classification.

Distinction in the structures properties reflects in distance from authors to a corpus pattern. Consider the mean pattern of the corpus by averaging feature vectors of the all texts and call it a lexicon of the language. The distances from the lexicon to the authors in L1 norm are shown that the 33 authors from previous section are located at distances from 0.11 to 0.19. It should be noted that the closeness to the lexicon of these authors is not determined by the number of texts. Although there are no authors with only 5 texts, not all of them have more than one hundred books. The distribution of the text amount in these authors is as follows: 16 have fewer than 30 texts, 12 have between 30 and 100 texts, and 5 have more than 100 texts. Thus, in half of the cases, the writers possess a relatively small number of texts. The patterns without connections (2,249 authors) have distances from 0.17 to 0.65 relative to the mean corpus. Thus, the authors with zero degree are shifted to the right compared to the patterns with stability large values of degree.

An influence of proximity the patterns and the text features with the lexicon to the recognition is considered below. Fig. 9 shows the dependency of the classification errors on distances from the authors to the lexicon (red dashed line). The distance values are plotted along the x-axis in increments of 0.01. The red solid line corresponds to the distance distribution. The authors are grouped by their distance's location in the increment intervals. For each group, the total number of author's texts is calculated and based on them, the number of incorrectly classified texts relative to the corpus. The group error is determined by division the number of wrong classifications on the total amount. Fig. 10 shows same distributions as in Fig. 9, but relative text-to-lexicon distances. The texts are grouped according to their location in the distance intervals. Classification errors were calculated for groups of texts. The constant values in both graphs are the result of combining intervals caused by a small number of texts and authors (and sometimes zero) into intervals with boundary values.

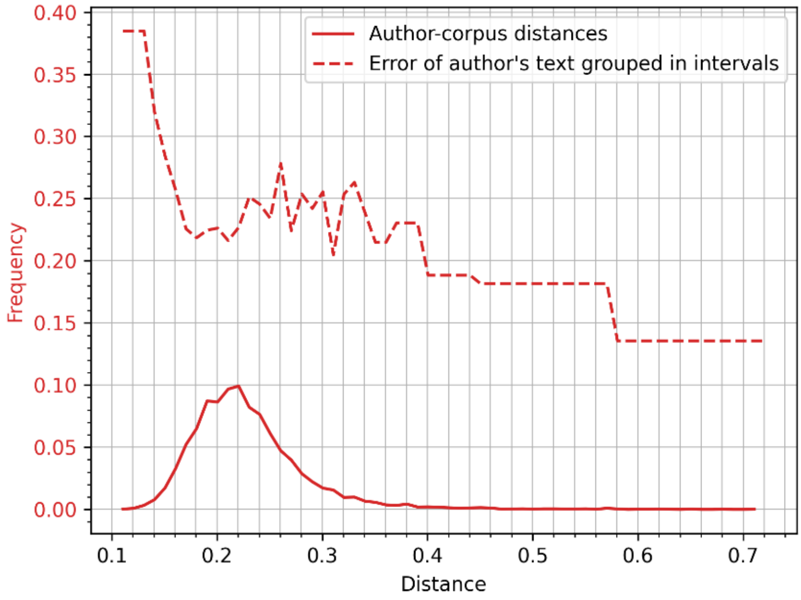

Fig. 9. Error distribution (red dashed line) grouped by intervals of distances from author to the corpus (red line).

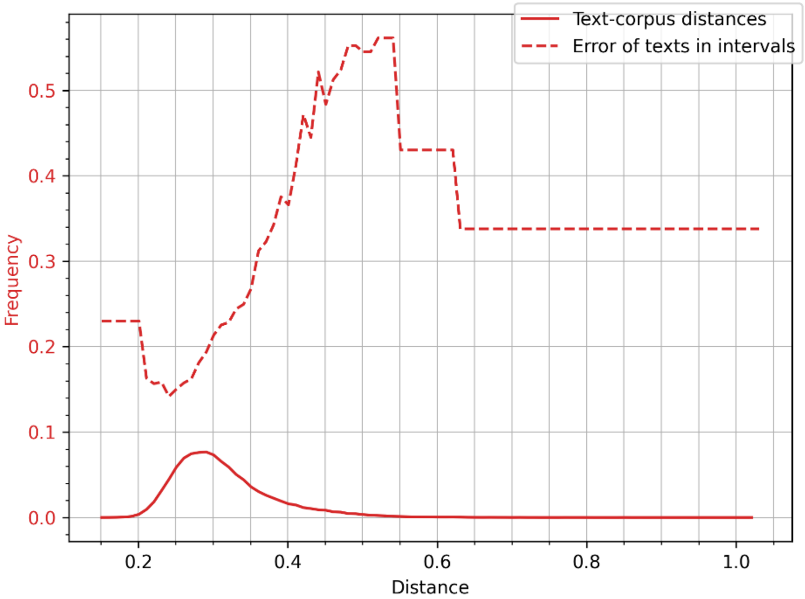

Fig. 10. Error distribution (red dashed line) grouped by intervals of distances from text to the corpus (red line).

Analysis of these two figures gives the following conclusions. Firstly, the distribution curves of the errors show different dependences in author and text cases. According to the error behavior in Fig.9, four intervals can be allocated: less 0.17, between 0.17 and 0.23, from 0.23 to 0.4 and greater 0.4. It can be seen, that authors, whose patterns are located at greater than 0.4, are recognized by the classification method with an error of about 13%. This is three times better than recognizing error of the authors with the less 0.17 distances. Consequently, the far of the lexicon authors are recognized on average better than patterns who are similar to the mean corpus. The main number of patterns is located at a distance from 0.17 to 0.23 (Fig. 9), and in this interval the error has a stable local minimum value of about 22%. Further, up to distances of 0.4, recognition has an oscillatory character from 20% to 27%, on average at the level of 24%.

A similar division can be made for the text distributions in Fig. 10. The four intervals have the following boundaries: less 0.23, from 0.23 to 0.35, between 0.35 and 0.54 and greater 0.54. As in the case of the authors, the error in first zone is less than in the last zone, but the difference isn’t so significant. More crucial changes of the error values are occurring in the third interval. In this interval, the error is inverse proportional to the distances. It can be assumed that the texts become far away not only from the mean corpus, but also from own pattern. The last interval consists of texts of authors, who located at large distances from the lexicon (the fourth group according to the distribution of authors). The major part of texts laying between 0.25 and 0.35, which contains a local minimum with the error of 14%. An appearance of the local minimums on the both graphs indicate a different influence on the errors. Before the local minimum, proximity of the authors or text to the lexicon increase mistakes in recognition, hence, texts and patterns too close (in mutual distances) to each other. After the local minimum, the influence of writing individuality on classification increases, and in the case of authors it provided better separation between patterns. In the text case, likely, texts far from the mean corpus are also far from own author.

Let's look at Fig. 9 again. There are 132 authors in total, whose patterns are located further than 0.4. Of these, almost half (63 authors) have zero degree in the graphs. Recall that a total of 2,249 authors with a zero degree of neighborhood were identified. Consequently, distant authors often have zero degree of neighborhood, whereas the converse is not true.

It is interesting to note that among the authors with a zero degree of neighborhood there are no "classical" writers. There are also completely no "serial" writers with more than 50 works. The next category of "famous writer" is rather arbitrary. Nevertheless, there are very few "famous" authors among these 2,249 – only 1%. The rest of the authors are so-called "self-publishing". This is the most common type of authors in this group.

As for the other half of the authors, who have a distance greater than 0.4, they are mostly world-famous philosophers: Aristotle, Herodotus, Plato, Plutarch, Nietzsche, Heidegger, Kant, Jung, Freud, Russell, Kierkegaard, Schopenhauer, Frank, Hegel, etc. At the same time, their degree of neighborhood is not zero. The text recognition error of this group varies from zero to 100%. In this group, 50 authors out of 132 or about 38% are recognized without errors. Note, that 2,191 out of 8,287 authors were unmistakably recognized, that is, 26%. Therefore, authors far from the center of the corpus are recognized with zero error more often than others.

The authors, which are close to the lexicon (distances less than 0.17) make up a completely different picture. These authors are mostly authors of popular series with several dozen works, for example: Roy Oleg, Robert Sheckley, Sandra Brown, Lynn Graham, et al. Almost all of them (27 out of 33) are the centers of clusters of neighbors of the order from 1 to 20. Consequently, the consistently high degree value of the nearest neighbor graphs mainly corresponds to the authors close to the average corpus pattern. In general, the opposite is not true.

5. Conclusions

The use of k-nearest neighbor graphs made it possible to identify four classes of patterns in which error models differ significantly. These classes are determined by the distance from the pattern to the corpus mean. The first class is the patterns close to the lexicon, located at a distance of 0.11-0.17. The second group includes the major amount of the authors and locates from 0.17 to 0.23. The third is the most complex in terms of error behavior, a fluctuating by the error values group of authors with distances from 0.23 to 0.40. The fourth group is the far tail of the distribution.

The method of visualizing author relationships within the corpus using a nearest-neighbor graph allowed the identification of certain stable structures. The neighbor graph of 1 to 20 neighborhood shows stable structures, which have the concentration of edges. The number of such points is sufficiently large, amounting to 33. The pattern in the points belong to the first group of distances to the lexicon. Moreover, this proximity is not attributed to a large number of texts or a high overall text volume. The recognition error for the texts of these authors is slightly less than 40%, that is higher than the corpus accuracy. Texts from authors who are nearest neighbors to one of the 33 patterns are, in nearly half of the misclassification cases, confused with texts from these same 33 authors. Authors, whose don't have incoming edges in any of the 20 nearest-neighbor graphs, are also exhibit large the recognition error. However, unlike the 33 authors, they are weakly confused with the same properties' patterns or with their nearest neighbors (from 1 to 20). These authors, who are not nearest to anyone, are located at a considerable distance from the corpus (second, third, and fourth groups). Thus, the properties of the graph nodes and their distances to the lexicon may indicate the recognizable characteristics of the texts by these authors. Logging such behavior may prove useful in developing heuristic methods to improve the classification performance of the k-nearest neighbor approach.

An example of author's patterns in the form of multidimensional n-gram distributions is interesting from the point of view of the approach to classification according to the most popular features of the text in the task of authorship attribution. Moreover, presented Russian Corpus of literary texts is one of the largest corpus of the Russian literary language, for comparison, the National Corpus of the Russian language contains about 11 thousand texts.

The use of nearest neighbor graphs has another important aspect related to the fact that for the sampling of independent random variables, the probabilities of the realization of certain structures are known quite accurately. Thus, estimating the probability of a relationship between the data based on the observed graphical structure allows us to choose and justify model choice in which such relationships can be explained.

Acknowledgments

This work is supported by RSCF, project № 23-71-10055

References

1. Jackson P., Benjamins I. Natural Language Processing for Online Applications Text Retrieval, Extraction and Categorization // Publishing Company Amsterdam, Philadelphia. 1984. 237 p.

2. Manning C., Schutze H. Foundation of Statistical Natural Language Processing: The MIT Press Cambridge. 1999. 606 p.

3. Koehn P. Statistical machine translation: Cambridge University Press. 2009. 433 p.

4. Vulic I., De Smet W., Tang J., Moens M.-F. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications // Information Processing & Management. 2015. Vol. 51, no. 1. Pp. 111–147.

5. Hirschberg J., Manning C. D. Advances in natural language processing. // Science. 2015. 349(6245). P.261-266.

6. Goldberg Y. A primer on neural network models for natural language processing // Journal of Artificial Intelligence Research. 2016. V. 57. P.345-420.

7. Sun S., Luo C., Chen J. A review of natural language processing techniques for opinion mining systems. // Information Fusion. 2017. 36. P.10-25.

8. Young T., Hazarika D., Poria S., Cambria E. Recent trends in deep learning based natural language processing. // IEEE Computational Intelligence Magazine. 2018. 13(3). P.55-75

9. Belov S., Zrelova D., Zrelov P., Korenkov V. Obzor metodov avtomaticheskoj obrabotki tekstov na estestvennom yazyke [Overview of methods for automatic natural language text processing] // System Analysis in Science and Education. 2020. № 3. Pp. 8–22. [in Russian]

10. Qiu X. et al. Pre-trained models for natural language processing: A survey // Science China Technological Sciences. 2020. T. 63. №. 10. C. 1872-1897.

11. Andrzejewski D., Buttler D. Latent topic feedback for information retrieval // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.KDD '11. 2011. Pp. 600–608.

12. Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. 2009. Pp. 27–34.

13. Balikas G., Amini M., Clausel M. On a topic model for sentences // Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16. USA, 2016. Pp. 921924.

14. Newman D., Noh Y., Talley E., Karimi S., Baldwin T. Evaluating topic models for digital libraries // Proceedings of the 10th annual Joint Conference on Digital libraries, JCDL '10. USA. 2010. Pp. 215–224.

15. Scherer M., von Landesberger T., Schreck T. Topic modeling for search and exploration in multivariate research data repositories //Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings 3. – Springer Berlin Heidelberg, 2013. – C. 370-373.

16. Chen G. H., Shah D. Explaining the success of nearest neighbor methods in prediction // Foundations and Trends in Machine Learning, 10(5-6). 2018. P. 337–588.

17. Hajebi K., Abbasi-Yadkori Y., and et. al. Fast Approximate Nearest Neighbor Search with k-Nearest Neighbor Graph. // Proc. of 22-d International Joint Conference on Artificial Intelligence, 2009. P. 1312-1317.

18. Raygorodski A. M. Modeli Interneta [Internet models]. Dolgoprudniy: «Intellekt», 2013. 64 p. [in Russian]

19. Leskovec J., Chakrabarti D. and et. al. Kronecker graphs: an approach to modeling networks // Machine Learning Research, 2010. V. 11. P. 985-1042.

20. Eppstein D., Paterson M., Yao F. On Nearest-Neighbor Graphs // Discrete & Computational Geometry, 1992, pp. 1-20.

21. Kislitsyna M., Orlov Yu. Statisticheskij analiz polnogo korpusa khudozhestvennoj literatury na russkom yazyke i raspoznavanie avtora [Statistical analysis of the complete corpus of fiction in Russian and recognition of the author] // Keldysh Institute preprints. 2024. № 17. P. 1-24. [in Russian]

22. Kislitsyn A.A., Orlov Yu.N., Goguev M.V. Investigation of the properties of first nearest neighbors' graphs // Scientific Visualization. 2023. V. 15. № 1. P. 17-28.

23. Kislitsyn A.A. Investigation of statistics of nearest neighbor graphs // Mathematical Models and Computer Simulations. 2023. V. 15. № 2. P. 235-244.

24. https://networkx.org/ (last date of request 07.04.2025)