Near-duplicate handwritten document detection without text recognition

Oleg Bakhteev
Antiplagiat, Dorodnicyn CC FRS CSC RAS
Moscow, Russia
bakhteev@ap-team.ru

Rita Kuznetsova
MIPT
Moscow, Russia
rita.kuznetsova@phystech.edu

Andrey Khazov
Antiplagiat
Moscow, Russia
khazov@ap-team.ru

Aleksandr Ogaltsov
Antiplagiat
Moscow, Russia
ogaltsov@ap-team.ru

Kamil Safin
MIPT
Moscow, Russia
kamil.safin@phystech.edu

Tatyana Gorlenko
Antiplagiat
Moscow, Russia
gorlenko@ap-team.ru

Marina Suvorova
Antiplagiat
Moscow, Russia
suvorova@ap-team.ru

Andrey Ivahnenko
Antiplagiat
Moscow, Russia
ivahnenko@ap-team.ru

Pavel Botov
Antiplagiat
Moscow, Russia
botov@ap-team.ru

Yury Chekhovich
Antiplagiat, Dorodnicyn CC FRS CSC RAS
Moscow, Russia
chehovich@ap-team.ru

Vadim Mottl
Dorodnicyn CC FRS CSC RAS
Moscow, Russia

Abstract

The paper presents a novel method for near-duplicate detection in handwritten document collections of school essays. A large amount of online resources with available academic essays currently makes it possible to cheat and reuse them during high school final exams. Despite the importance of the problem, at the moment there is no automatic method for near-duplicate detection for handwritten documents, such as school essays. The school essay is represented as a sequence of scanned images of handwritten essay text. Despite advances in recognition of handwritten printed text, the use of these methods for the current task is a challenge. The proposed method of near-duplicate detection does not require detailed markup text, which makes it possible to use it in a large number of tasks related to the information extraction in zero-shot regime, i.e. without any specific resources written in the processed language. The paper presents a method based on series analysis. The image is segmented into words. The text is characterized by a sequence of features, which are invariant to the author's writing style: normalized lengths of the segmented words. These features can be used for both handwritten and machine-readable texts. The computational experiment is conducted on IAM dataset of English handwritten texts and the dataset of real images of handwritten school essays.

Keywords: handwritten text analysis, near-duplicate detection, word segmentation, time series analysis
DOI: 10.28995/2075-7182-2021-20-47-57

Поиск почти дубликатов рукописных текстов без распознавания текста

Бахтеев О. Ю.
Антиплагиат, ФИЦ ИУ РАН
Москва, Россия
bakhteev@ap-team.ru

Кузнецова Р. В.
МФТИ
Москва, Россия
rita.kuznetsova@phystech.edu

Хазов А. В.
Антиплагиат
Москва, Россия
khazov@ap-team.ru

Огальцов А. В.
Антиплагиат
Москва, Россия
ogaltsov@ap-team.ru

Сафин К. Ф.
МФТИ
Москва, Россия
kamil.safin@phystech.edu

Горленко Т. А.
Антиплагиат
Москва, Россия
gorlenko@ap-team.ru

Суворова М. А.
Антиплагиат
Москва, Россия
suvorova@ap-team.ru

Ивахненко А. А.
Антиплагиат
Москва, Россия
ivahnenko@ap-team.ru

Ботов П. В.
Антиплагиат
Москва, Россия
botov@ap-team.ru

Чехович Ю. В.
Антиплагиат, ФИЦ ИУ РАН
Москва, Россия
chehovich@ap-team.ru

Моттль В. В.
ФИЦ ИУ РАН
Москва, Россия

Аннотация

Рассматривается задача поиска почти-дубликатов в коллекции сканированных изображений школьных сочинениях. Сочинение представляется набором изображений рукописного текста, написанного автором. Актуальность задачи обусловлена наличием больших библиотек школьных сочинений, которые могут использоваться школьниками в качестве источника заимствования при написании собственного сочинения. На текущий момент не существует автоматических методов анализа сочинений на наличие заимствований. Несмотря на успехи в области распознавания рукописного текста,применение данных методов для рассмотренной задачи затруднительно. Для решения задачи предлагается рассматривать текст, находящийся в изображении, как последовательность. Предлагается метод, заключающийся в сегментации слов в изображении. Текст характеризуется последовательностью признаков, полученных на основе сегментации. В качестве такого признака выступает нормализованная длина извлеченных из изображения слов. Полученные статистики являются инвариантными по отношению к почерку автора, а также могут использоваться как для рукописных, так и для машиночитаемых текстов. Предложенный метод поиска почти-дубликатов не требует наличия аннотированных корпусов изображений, и потому может быть применим для низкоресурсных языков. Для подтверждения работоспособности метода проводятся эксперименты на англоязычном корпусе IAM, а также выборке реальных изображений рукописных текстов школьных сочинений.

Ключевые слова: рукописные изображения, поиск почти-дубликатов, сегментация слов, анализ временных рядов

1 Introduction

The paper is devoted to the analysis of academic essays and textual reuse detection in them. We consider the problem on the example of school essays written during high school final exams in Russia. A standardized system of assessment for the essay makes it possible to reuse some text or parts of the texts from open collections of school essays available on the Internet. We refer to the problem as near-duplicate detection, but not plagiarism detection problem because the proposed method is robust to slight changes in compared documents. Also, near-duplicate detection is a more precise formulation since plagiarism is a fact that is approved by experts after a detailed analysis of near-duplicate passages.

The main feature that makes the problem hard to analyze is Russian cursive which is really variable in terms of styles of writing letters and connecting them with each other. This is a known feature, but even now datasets for Russian handwritten text recognition are proprietary and not available for the public. Therefore we can't use state-of-the-art handwritten recognition algorithms due to the lack of datasets to train on. Since word recognition is a crucial component for subsequent text reuse detection it is not possible to obtain decent detection quality.

Our contributions are:

- we propose a novel method that avoids the stage of word recognition and directly applicable to the image of the analyzed document;
- we present the dataset of real handwritten school essays and baseline for the text reuse detection task without word recognition;
- we compare the performance of our method with state-of-the-art recognition-based algorithm using IAM dataset for handwritten text recognition in English.

2 Related Work

The problem of finding sources of textual reuse in academic essays is a challenge and can be considered as critical for the educational system [13, 8, 19]. Despite the probably massive nature of the problem, at the moment there is no automatic method of text reuse detection in school essays. The closest work in this area [15] involves an automated system for collecting and analysing essays written in English. The methods described in it are not directly applicable to our problem since the student works considered in [15] are written using printed letters, which are simpler to analyse. The method to compare two document images is presented in [5]. It is based on a similarity measure on the top of word bounding boxes vector representations that are obtained by convolution neural network. The authors pointed out the problem of low data resources, but they deal with it by generating synthetic data and perform transfer learning on IAM dataset. In contrast, we have a slightly different task of comparison handwritten documents with a collection on properly printed documents. Also, we propose not to generate additional data, but perform in zero-shot manner without any learning.

The major works in the area of handwritten text analysis are based on the text recognition methods [3, 15, 18, 12]. Currently, the methods based on deep learning achieve rather good performance on the handwritten recognition task [3, 18], which potentially makes it possible to use it with a combination of modern plagiarism detection systems [2]. The main disadvantage of such methods is the requirement for the presence of markup: to optimize the parameters of recognition models, a significant corpus of annotated texts is required. Therefore this method is not applicable if the documents are written in the language with a lack of such markup. An example of such language is Russian: despite the amount of works devoted to handwritten text and distinct characters recognition [6, 11], to the best of our knowledge currently there is no available publicly annotated corpus for the Russian language.

This paper presents a simple yet efficient method for near-duplicate school essay detection. The method is based on the word segmentation with further analysis of word lengths extracted from the texts. The word segmentation is a well-studied problem, which can be conducted much simpler than word recognition and basically does not require any markup, therefore the proposed method can be applied as a zero-shot method for languages without any annotated corpus for recognition model training. We analyse the lengths of words extracted from the texts and empirically show that they can be considered as features for the information retrieval algorithms invariant to the author's writing style. For similar text detection we employ different methods of time series and sequence alignment [4, 14]. The computational experiment is conducted on two datasets of scanned images: IAM dataset of handwritten texts [9] and the real dataset of the images of handwritten school essays.

3 Problem statement

Consider the problem of near-duplicate detection as an information retrieval problem. Given a dataset of school essays, which are represented by scanned images of handwritten text:

$$D_{\text{susp}} = \{d^i_{\text{susp}}\}.$$

There is also given a collection of documents $D = \{d^j\}$, which can be represented both as scanned images or text in the machine-readable format. We suppose that for each essay $d^i_{\text{susp}} \in D_{\text{susp}}$ there is only one document in collection, which was employed as a source of text reuse:

$$g : D_{\text{susp}} \to D.$$

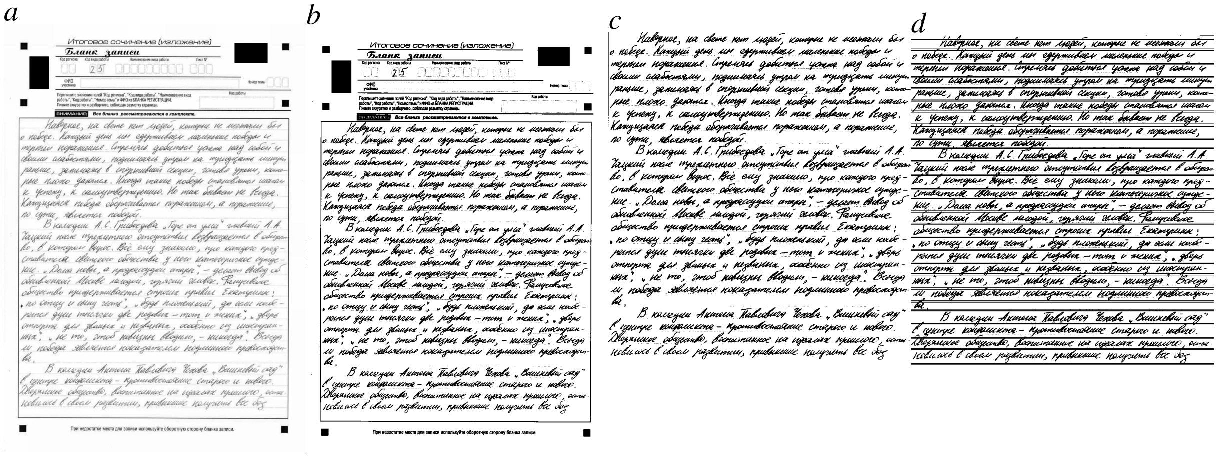

Figure 1: School essay page example: *a)* original image; *b)* binarized image; *c)* text area segmentation; *d)* line segmentation.

The major quality criterion for this task is Recall@K maximization, where Recall@K is a ratio of relevant documents in the most similar K documents retrieved by our method:

$$\text{Recall@}K = \frac{1}{|D_{susp}|} \sum_{d^i_{susp}} |f(d^i_{susp})\text{@}K \cap \{g(d^i_{susp})\}|, \quad (1)$$

where f is a document retrieval model, $f(d^i_{susp})\text{@}K$ is a set of top-K documents the most similar to the document d^i_{susp}.

After the model found a probable text reuse source for the suspicious document, the source should be verified by the expert. In practice the expert can analyse only a small number of retrieved documents, therefore the formal optimization task is to find a mapping, that maximizes Recall@1 for our dataset:

$$\hat{f} = \arg\max_{f \in \mathcal{F}} (\text{Recall@1}(f, g, D, D_{\text{susp}})),$$

where $\mathcal{F}$ is a family of considered retrieval models.

3.1 Near-duplicate detection using word segmentation

The proposed method is based on the considering the text as a series of features [10]. We propose to segment the document into words without its further recognition. The extracted words are further transformed into features. In this paper we use only the word length as such feature, however other features, such as word height or number of ligatures can potentially improve current near-duplicate detection quality. Opposite to challenging text recognition problem these features are rather simple to extract from the handwritten document.

The school essay d^i_{susp} is represented as a sequence of scanned images of handwritten text. The essay form is standardized and has clear ruled lines, therefore the problem of line segmentation can be solved using a rule-based line segmentation algorithm. The image preprocessing consists of image binarization, text area extraction and line segmentation. The example of preprocessing steps application is shown in Figure 1.

The further image analysis step is word segmentation. For this problem, we use the method based on connectivity component analysis [7]. To take into account word cursive during word length analysis we use a deslatning algorithm similar to [17]. This step is significant especially if the essay author uses significant letter tilt and also helps to segment words more accurately. After that, we extract connectivity components and determine the thresholds for spaces between words and between characters. Since these thresholds depend on the author writing style we determine them dynamically using a Gaussian mixture with 2 components:

$$s \sim \alpha \mathcal{N}(m_1, s_1) + (1-\alpha)\mathcal{N}(m_2, s_2),$$

where s is a distance between connectivity components, $\alpha \in (0, 1)$. We suppose that the component with a smaller mean corresponds to the space distance between characters in words. We unite the connectivity components with distance between that is more likely to correspond to this component. The example of word segmentation is shown in Figure 2. The list of numbers presented in Figure 2.e is a list of lengths in pixels of the bounding boxes of the extracted words. We normalize them by dividing by the average box length extracted from the text. The resulting sequence of normalized word length is shown in Figure 2.f.

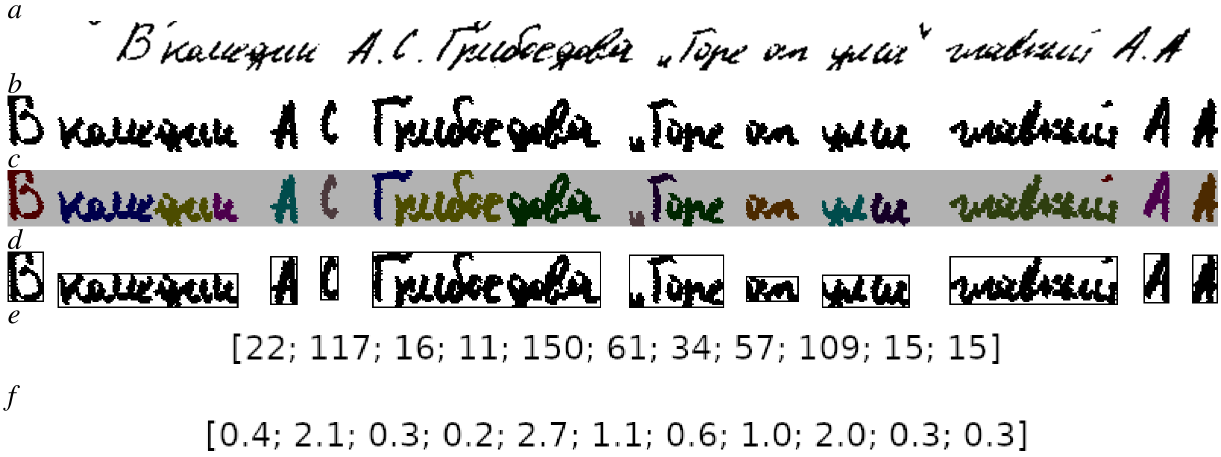

Figure 2: Word segmentation example: *a)* original image line; *b)* deslanting; *c)* connectivity component extraction; *d)* word segmentation result; *e)* word length extraction; *f)* word length normalizing.

After the word segmentation for each image we obtain a sequence of lengths extracted from the image. We normalize this sequence by average extracted word length and consider it as a feature that characterizes the essay. For the essay comparison we employ functions based on the dynamic time warping method [4]:

$$\text{DTW}(x^1, x^2) = \text{dtw}_{t_1, t_2},$$

$$\text{dtw}_{i,j} = ||x^1_i - x^2_j||_2 + \min(\text{dtw}_{i,j-1}, \text{dtw}_{i-1,j-1}, \text{dtw}_{i-1,j}),$$

where x^1, x^2 are the sequences of lengths t_1 and t_2 correspondingly.

The computational complexity of DTW which is $O(t_1 \cdot t_2)$. In this paper we employed DTW function and its modification FastDTW [14], which has linear computational complexity. Although these methods are well-known for our knowledge there is no research of usage such representation for near-duplicates detection of handwritten texts. We are inspired by the work [10], which shows that considering text as time series and subsequent outlier detection is a fruitful approach to the problem of intrinsic plagiarism detection.

4 Experiment

In order to demonstrate the performance of the proposed method we conducted computational experiments with two image datasets: IAM dataset [9] and a dataset of real school essays [1]. The brief statistics about the used dataset is represented in Table 1. For better experiment reproducibility we used two datasets as a document collection D for the Russian language: a collection of essays mined from the Internet and Taiga corpus [16]. To the best of our knowledge currently there is no available publicly available

[1] The dataset is available at http://bit.ly/ap_handwritten

Table 1: Statistic about the used dataset

Suspicious documents, D_{susp}			
Dataset	Language	Document number	Average word number
IAM [9]	English	336	76
School essays	Russian	89	263
Document collection, D			
Dataset	Language	Document number	Average word number
IAM [9]	English	992	75
School essays	Russian	17361	503
Taiga[16]	Russian	15197	287

dataset annotated for handwritten text recognition, therefore we used IAM dataset for comparison with the text recognition-based model.

As a quality criteria for both experiments we used recall function (1): Recall@1, Recall@10 and Recall@100.

Experiment on IAM dataset. The dataset consists of handwritten images of text segmented into lines. Each line has an annotation file with information about each word in line. The dataset is split into *Train, Test* and *Validation* parts. We used the images from the *Test* split as a set of suspicious documents $D_{\text{susp}}, |D_{\text{susp}}| = 336$ and all the text documents from the dataset as a collection of documents $D, |D| = 992$. We used *Train* part of dataset to tune hyperparameters of the proposed algorithm. Some of the images of the dataset contains identical texts written by different authors. We did not use this information and considered all the images as independent objects.

This dataset was used to compare the proposed method with text recognition-based models. For the comparison we used a model from [3], a neural network-based model achieving state-of-the-art results on multiple handwritten text recognition datasets. We trained the model with different percentages of the images from *Train* subset: $\{10\%, 20\%, 50\%, 100\%\}$. The performance of these models is presented in Table 2. For each percentage we ran the training procedure 5 times for 1000 epochs, the results were averaged.

We evaluated the word segmentation algorithm used in the proposed method using the methodology described in [1]. For the used word segmentation algorithm we got *Precision*=0.8, *Recall*=0.7, F_1=0.75, which is quite comparable to other word extraction algorithms.

As a distance function between the documents we used a cosine distance between the collection document and text extracted from the image:

$$\cos(\mathbf{v}_1, \mathbf{v}_2) = \frac{\langle \mathbf{v}_1, \mathbf{v}_2 \rangle}{||\mathbf{v}_1||_2 \cdot ||\mathbf{v}_2||_2},$$

where $\mathbf{v}_1, \mathbf{v}_2$ are the bag-of-words vectors of the texts of the compared documents.

For the methods based on series analysis we filtered one-character words from the document collection D. We found that this heuristic gives a significant performance improve since the large amount of “a” words in texts lowers the chances to correctly align short texts. The results for the experiment are shown in Table 2. The *Ground Truth Word lengths* method corresponds to the application of series analysis to the word lengths from the dataset annotation. The results for this method show a performance that can be potentially achieved if the word segmentation method works perfectly without any error. The results show that the performance of the text recognition-based model dramatically decreases with size of the training dataset. As we can see, the proposed method performance is comparable with state-of-the-art recognition method that is trained on half of the dataset, however the proposed algorithm achieves comparable performance in zero-shot manner. It can be used for languages with a lack of ground-truth data for handwritten word recognition which is actually very frequent case.

Table 2: Word error rates (WER) for the recognition-based models

Method	WER
[3], 10% of the dataset used for training	0.921 ± 0.001
[3], 20% of the dataset used for training	0.836 ± 0.010
[3], 50% of the dataset used for training	0.546 ± 0.027
[3], 100% of the dataset used for training	0.187 ± 0.000

Table 3: Experiment results for the IAM dataset.

Method	Recall@1	Recall@10	Recall@100
[3], 10% of the dataset used for training	0.00 ± 0.00	0.04 ± 0.01	0.23 ± 0.03
[3], 20% of the dataset used for training	0.02 ± 0.01	0.15 ± 0.04	0.47 ± 0.04
[3], 50% of the dataset used for training	0.74 ± 0.05	0.89 ± 0.03	0.98 ± 0.01
[3], 100% of the dataset used for training	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Proposed, DTW	0.66	0.78	0.89
Proposed, FastDTW	0.55	0.69	0.82
Ground Truth Word lengths, DTW	0.97	0.99	0.99
Ground Truth Word lengths, FastDTW	0.92	0.95	0.97

Experiment on the dataset of real handwritten school essays. The dataset of suspicious essays D_{susp} consists of 89 images of school essays. For each image we have the corresponding text without per-word annotation. We split the dataset into two parts: 18 images for *Train* part and 71 images for *Test* part. *Train* part was used to tune hyperparameters of the proposed algorithm.

As a collection of texts D we used two different datasets. The first dataset is a dataset of school essays mined from the Internet. The dataset consists of 17361 documents. In order to increase the reproducibility of the experiment we also used the second dataset, which was constructed as a subset of Taiga corpus [16]. We used a subset of *proza.ru* texts included in this corpus. We used only texts from the year 2009, which length is similar to typical essay length: from 150 to 400 words. The final collection size was 15197 texts. Both the datasets does not contain the real sources of the suspicious documents D_{susp}. For each document d^i_{susp} we also added into collection D the real source of the document, thus during the experiment there is only 1 real source document in the collection D.

The results for the experiment are shown in the Table 4, Table 5. As we can see the proposed method gives rather good results for both collections. For the Taiga corpus we also estimated the time for one school essay processing. All the experiments were run on the computer with 16 GB RAM and Intel Core i5 CPU. For both the experiments we used only one core. As we can see, FastDTW performs significantly faster, however, the quality of the proposed method with DTW is better. One of the further directions in the development of the proposed method is the combination of these functions in order to obtain a trade-off between the quality and speed of the method.

We also analyzed the dependence of the proposed method on the essay length and its similarity to the original texts. Firstly, we conducted an experiment truncating all the analyzed sequences, extracted from the collection D and suspicious documents D_{susp}. We considered different truncation percentage: from 10% to 90%. For the experiment DTW function was used. The document collection D is Taiga corpus. The results are shown in Figure 3. As we can see, the proposed method works poorly on the small texts, which also can explain the difference in performance on the IAM dataset and the dataset of the school essays: the average essay length is much longer than the average document from the IAM dataset, 65 words after removing short words in IAM dataset versus 257 words in essays.

Secondly, we analyzed the performance of the proposed method for the case, when the original text and suspicious document are partially different. For this experiment instead of including into the document

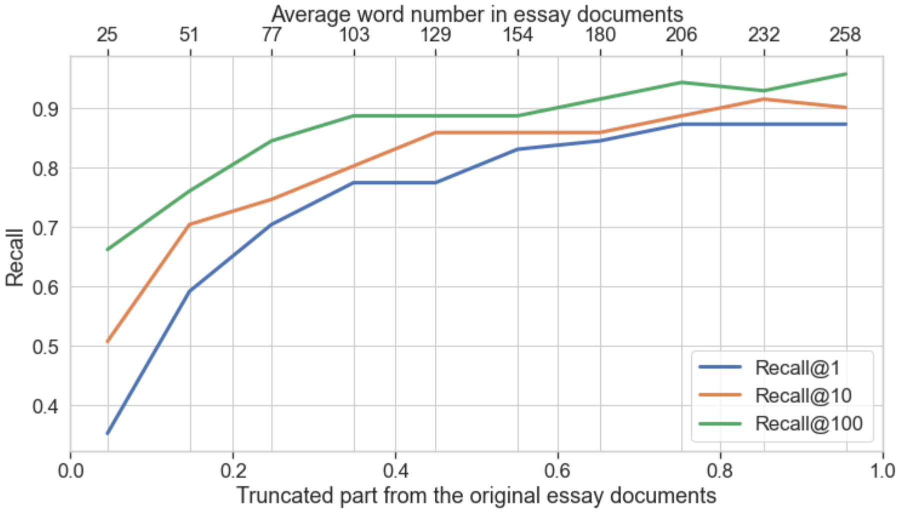

Figure 3: The dependency of the performance of the proposed method on the analyzed sequences lengths.

Table 4: Experiment results for the the dataset of real handwritten school essays using documents mined from the Internet as a document collection D.

Method	**Recall@1**	**Recall@10**	**Recall@100**
Proposed, DTW	0.93	0.99	1.0
Proposed, FastDTW	0.80	0.91	0.99

collection D the original essay text we randomly mixed it with another document from the collection. We considered different percentage p of the original essay text for this procedure. More formally, we conducted the following steps:

1. select the original essay text, extract word length sequence from this text;
2. randomly select subsequence of the sequence with $p\%$ of the original sequence;
3. randomly select document d from the collection D, extract word length sequence from this text;
4. randomly select subsequence of the collection document text series with $(100 - p)\%$ of the original sequence;
5. randomly insert the subsequence of the essay text into the subsequence of the collection document;
6. add the resulting subsequence into the series of the collection D;
7. remove the sequence of the document d from the collection D.

This algorithm simulates the situation, when the text was copied from the origin partially, with $p\%$ of text reuse. For this experiment we mix the original text with one of the documents d from the collection D, therefore the number of ground-truth source documents increases. We believe that this differs from a real-world setting, when the student often copies the text only from one origin. Therefore we remove the series of the document d from the collection D in order to have only one ground-truth origin for each essay. We considered different mixture percentage: from 70% to 90%. The experiment was run 5 times, the results were averaged. As for the previous experiment, we used DTW function and Taiga corpus for the document collection D. The results are shown in Figure 4

For further analysis we collected 25 essay images that use one text as a source of reuse. We built an alignment matrix [4] between them to demonstrate the proposed method operability. The matrix for these texts is shown in Figure 5a. In comparison, we also built alignment matrices between these essay images and random school essay texts. The result is shown in Figure 5b. The alignment matrix for the essay image and true text reuse source is strongly diagonal while the matrix between random texts does not demonstrate this matrix property.

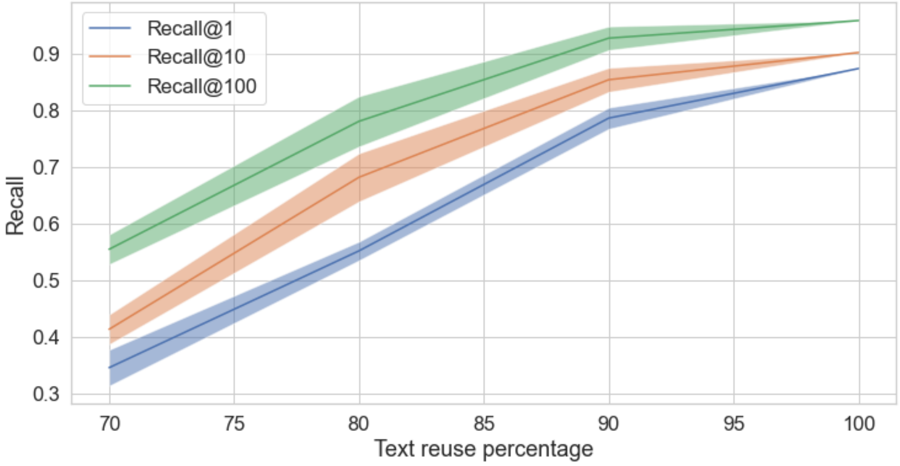

Figure 4: The dependency of the performance of the proposed method on the similarity between suspicious document and the original collection document. The results are averaged between different experiment runs.

Table 5: Experiment results for the the dataset of real handwritten school essays using subset of Taiga corpus as a document collection D.

Method	**Recall@1**	**Recall@10**	**Recall@100**	**Time per one essay, sec**
Proposed, DTW	0.87	0.90	0.96	73.4 ± 13.1
Proposed, FastDTW	0.66	0.70	0.79	3.9 ± 0.2

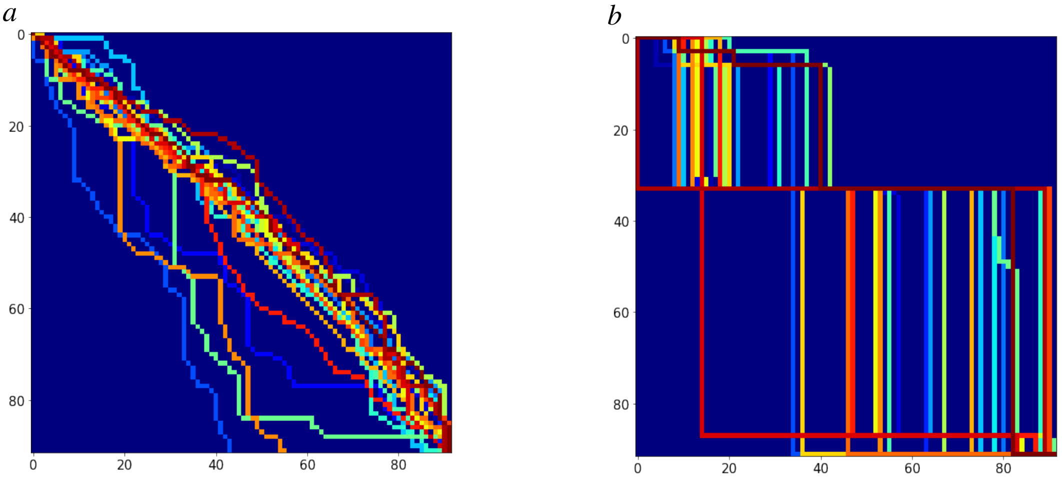

Figure 5: Alignment for image documents: *a)* with real text reuse source; *b)* with random documents.

To conclude the proposed method showed rather good quality for near-duplicate handwritten document retrieval. The method has a performance comparable to the performance of the text recognition-based methods and can be especially useful for low-resource languages that have no markup for recognition model training.

5 Conclusion

The paper is devoted to the problem of near-duplicate detection in handwritten school essay collections. The proposed method is based on word segmentation and further analysis of the extracted word lengths.

As a distance function between the essays, we analysed functions based on the dynamic time warping function. The computational experiment showed that the proposed method can efficiently work on large collections of school essays and comparable to the state-of-the-art handwritten text recognition methods. The future work includes analysis of different similarity functions and usage of different features that can be extracted from the text without its recognition.

Acknowledgements

This research is funded by RFBR, grant 19-29-14100.

References

[1] Axler Gregory, Wolf Lior. Toward a dataset-agnostic word segmentation method // 2018 25th IEEE International Conference on Image Processing (ICIP) / IEEE. — 2018. — P. 2635–2639.

[2] Discovering text reuse in large collections of documents: A study of theses in history sciences / Anton Khritankov, Pavel Botov, Nikolay Surovenko et al. // 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT) / IEEE. — 2015. — P. 26–32.

[3] End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network / Denis Coquenet, Clément Chatelain, Thierry Paquet, Ioan Grozny // arXiv preprint arXiv:2012.03868. — 2020.

[4] Giorgino Toni et al. Computing and visualizing dynamic time warping alignments in R: the dtw package // Journal of statistical Software. — 2009. — Vol. 31, no. 7. — P. 1–24.

[5] Krishnan Praveen, Jawahar C. V. Matching handwritten document images // European Conference on Computer Vision / Springer. — 2016. — P. 766–782.

[6] Liepieshov Kostiantyn, Dobosevych Oles. On recognition of Cyrillic Text // Workshop on Document Intelligence at NeurIPS 2019. — 2019.

[7] Louloudis Georgios, Gatos Basilios et al. Text line and word segmentation of handwritten documents // Pattern recognition. — 2009. — Vol. 42, no. 12. — P. 3169–3183.

[8] Ma Hongyan Jane, Wan Guofang, Lu Eric Yong. Digital cheating and plagiarism in schools // Theory Into Practice. — 2008. — Vol. 47, no. 3. — P. 197–203.

[9] Marti Urs-Viktor, Bunke Horst. The IAM-database: an English sentence database for offline handwriting recognition // International Journal on Document Analysis and Recognition. — 2002. — Vol. 5, no. 1. — P. 39–46.

[10] Methods for Intrinsic Plagiarism Detection and Author Diarization. / Mikhail Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, Vadim Strijov // CLEF (Working Notes). — 2016. — P. 912–919.

[11] Mustakimova Elmira. Offline Recognition of Russian Handwriting : Master's thesis / Elmira Mustakimova. — 2016. — Master thesis.

[12] Pandey Om, Gupta Ishan, Mishra Bhabani S. P. A Robust Approach to Plagiarism Detection in Handwritten Documents // International Symposium on Visual Computing / Springer. — 2020. — P. 682–693.

[13] Prevalence of Plagiarism among Medical Students / Vedran Frković, Josip Ažman, Tamara Turk et al. // Book of Abstracts. ZIMS 4. — P. 38–38.

[14] Salvador Stan, Chan Philip. Toward accurate dynamic time warping in linear time and space // Intelligent Data Analysis. — 2007. — Vol. 11, no. 5. — P. 561–580.

[15] Scaling Handwritten Student Assessments With a Document Image Workflow System / Vijay Rowtula, Varun Bhargavan, Mohan Kumar, C. V. Jawahar // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. — 2018. — P. 2307–2314.

[16] Shavrina Tatiana, Shapovalova Olga. To the methodology of corpus construction for machine learning:“Taiga” syntax tree corpus and parser // Proceedings of “CORPORA-2017” International Conference. — 2017. — P. 78–84.

[17] Vinciarelli Alessandro, Luettin Juergen. A new normalization technique for cursive handwritten words // Pattern recognition letters. — 2001. — Vol. 22, no. 9. — P. 1043–1050.

[18] Voigtlaender Paul, Doetsch Patrick, Ney Hermann. Handwriting recognition with large multidimensional long short-term memory recurrent neural networks // 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR) / IEEE. — 2016. — P. 228–233.

[19] Wrigley Stuart. Avoiding ‘de-plagiarism’: Exploring the affordances of handwriting in the essay-writing process // Active Learning in Higher Education. — 2019. — Vol. 20, no. 2. — P. 167–179.