**Методы и технологии анализа текстовых**

**штампов и именованных сущностей на основе массивов выпускных сочинений**

Данный проект посвящен анализу школьных сочинений и выявлению различных штампов внутри них. Актуальность проекта обусловлена наличием больших коллекций типовых школьных сочинений, которые потенциально могут быть использованы школьниками в качестве источников заимствования при написании собственного сочинения. Актуальность проблемы подтверждается рядом научных публикаций, посвященных анализу нарушений при написании сочинений, как зарубежных, так и российских авторов. Тем не менее, на данный момент не существует автоматических методов анализа сочинений на наличие корректных и

некорректных цитирований, штампов, а также заимствований. В первую очередь выполнение работ по данному проекту полезно проверяющим школьные работы экспертам, а также учителям при анализе и проверке домашних заданий в форме сочинений.

При разработке алгоритма поиска текстовых штампов в школьных сочинениях мы

столкнулись с проблемой: классические подходы к поиску штампов и заимствований направлены на анализ непосредственного текста сочинения. В то же время, в большинстве случаев, школьные сочинения пишутся учеником от руки, а затем сканируются или фотографируются. Это вносит дополнительное требования на разрабатываемый метод: метод поиска должен работать не только с документами, содержащими текст сочинения, но и с изображениями, являющимися сканами сочинений, написанных от руки. Пример такого скана представлен на Рис. 1. Несмотря на успехи в области распознавания печатного текста, применение данных методов для рассмотренной задачи затруднительно. Основной проблемой для подобных методов поиска является невозможность адаптироваться к большому числу вариантов почерка. Поскольку сочинения пишутся разными людьми, то дообучение системы распознавания на конкретный почерк автора отсутствует, что также усложняет задачу.

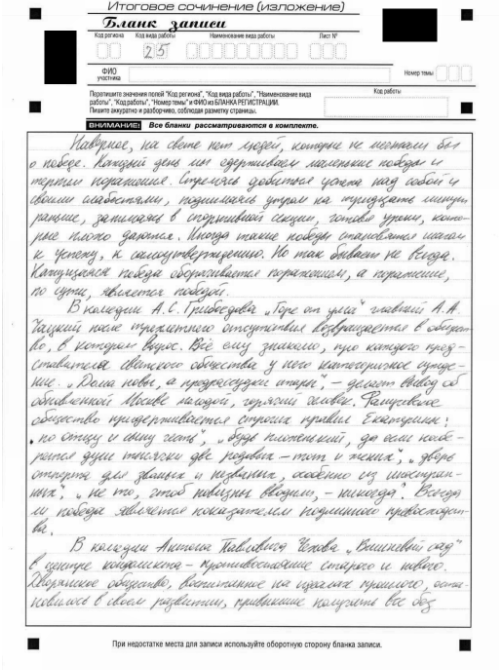


Рис 1. Пример анализируемого школьного сочинения.

Для решения рассматриваемой в проекте задачи было предложено сопоставлять тексту

числовую последовательность для дальнейшего сравнения полученных последовательностей.

Метод поиска заключался в выделении слов в изображении с дальнейшим извлечением

графических признаков. При таком подходе анализируемый текст характеризуется

последовательностью признаков. Получение этих признаков значительно проще, чем

распознавание самих слов в изображении. В рамках проведенного эксперимента тексту

сопоставлялись нормализованные длины извлеченных из изображения слов.

Полученная последовательность является инвариантной по отношению к почерку автора, а

также может использоваться как для рукописных, так и для машиночитаемых текстов. Для

сопоставления полученных последовательностей между собой рассматривался набор

методов выравнивания последовательностей и временных рядов. Схема работы разработанного метода представлена на Рис. 2.

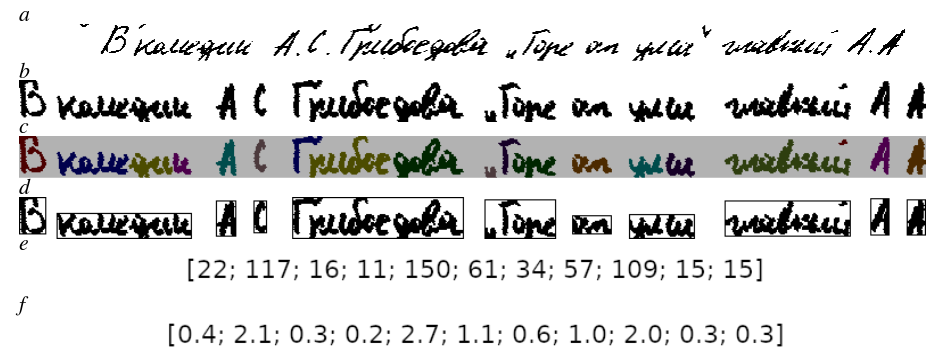


Рис 2. Схема работы разработанного метода: исходный текст (a) проходит нормализацию наклона почерка (b), затем из текста извлекаются связные компоненты (с), производится извлечение слов без распознавания (d). Для полученных слов извлекаются длины (e), производится их нормализация для дальнейшего сравнения (f).

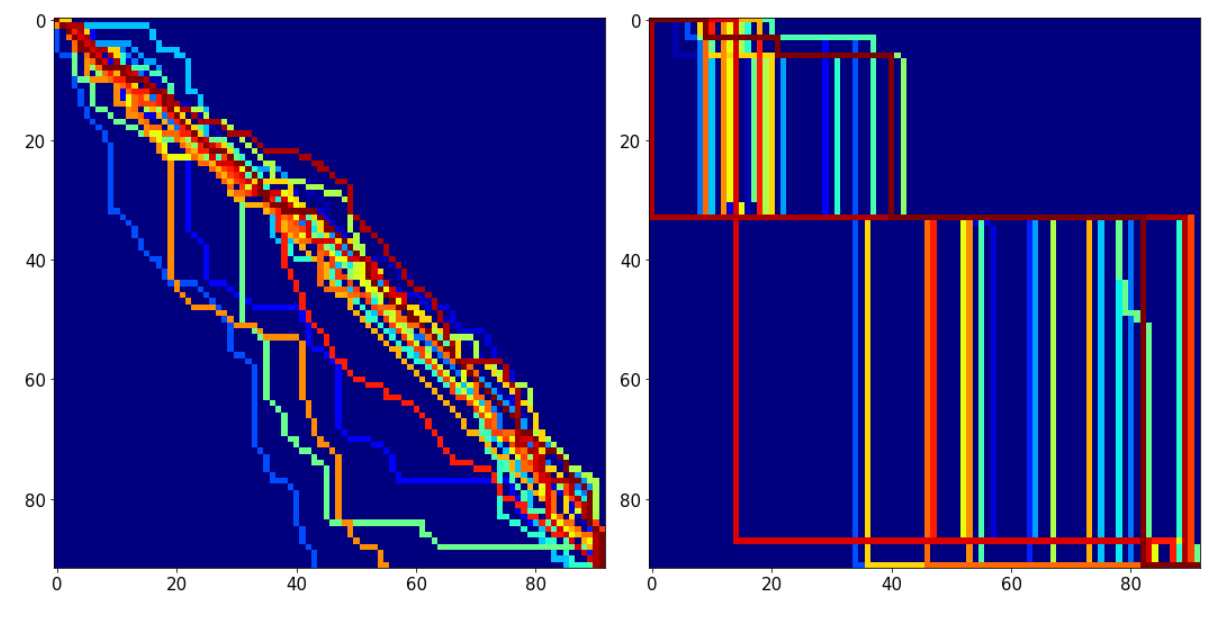


Рис. 3. Визуализация выравнивания между сканами одного и того же

сочинения, написанного разными почерками (слева) и случайными сочинениями (справа).

Видно, что матрица выравнивания одного и того же сочинения диагональна, в то время как

для случайных сочинений такое свойство не наблюдается.

В рамках проекта был разработан алгоритм кластеризации штампов с использованием предложенных методов. Пример анализа кластеризации текста в сочинениях приведен на Рис. 3. Был проведен анализ возможности применения разработанного метода для поиска совпадающих текстовых фрагментов различной длины. В рамках экспериментов было показано, что разработанный алгоритм позволяет находить совпадающие текстовые фрагменты в сочинениях небольшой длины, от ста слов. В то же время длина использованного фрагмента должна занимать значительную часть сочинения: не менее 75%.

В рамках последнего года исследований были проведены дополнительные исследования: в 2021 появилась работа, подтверждающая возможность обучения модели оптического распознавания текста на изображениях рукописного текста на русском языке. В то же время, обучающая выборка, которая применялась в этой работе, состоит из достаточно высококачественных фотографий текста с минимальным набором помарок. Такой набор подходит для обучения моделей распознавания текста, но не дает представления об обобщающей способности модели, и в частности, не отвечает на вопрос о том, насколько такая модель будет хорошо работать на сканах школьных сочинений, предположительно обладающих большим количеством помарок. Кроме того, далеко не всегда школьные сочинения сканируются с применением высококачественной аппаратурой: зачастую изображения получаются с помощью фотоаппаратов или камер мобильных устройств. Для исследования качества модели оптического распознавания текста и сопоставления с разработанным методом, была собрана небольшая выборка рукописных текстов, написанных в разных условиях фотографирования, в том числе с разной освещенностью и качеством фотографирования. Пример фотографии из данной выборки представлен на Рис. 4.

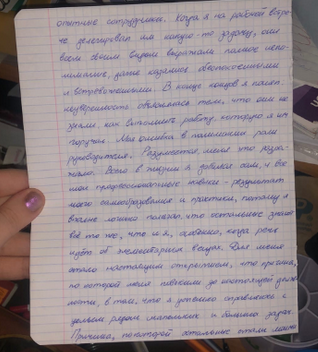


Рис 4. Пример фотографии сочинения из полученной выборки. В отличие от первой выборки, полученной в рамках работы над проектом, во второй выборке собраны фотографии сочинений, с разным уровнем освещенности, возможным наличием различных артефактов и фона.

Результаты экспериментов показывают, что обе модели работают хорошо на первой выборке, состоящей из сканов сочинений. В то же время, модель на основе оптического распознавания образов обладает большей обобщающей способностью и лучше справляется с изображениями низкого качества. Полученный эксперимент позволил определить границы применимости нашего метода, а также сделать вывод о возможности дальнейшего комбинирования различных моделей: модель распознавания текста обладает более высоким качеством, но в то же время, требует значительно больше вычислительных ресурсов. Поэтому возможно применение схемы, при которой сначала задействуется легковесный предложенный метод, а для более детального исследования текстовых штампов используется модель распознавания.