

Методы поиска почти-дубликатов рукописных документов в больших коллекциях текстов

Бахтеев Олег Грабовой Андрей Каприелова Мариам
Кильдяков Александр Сейил Темирлан Финогеев Евгений
Чехович Юрий

Антиплагиат

Проблема

- Во многих отраслях все еще присутствуют рукописные тексты, и множество таких текстов растет.
- На текущий момент не существует методов распознавания рукописного текста, работающих с достаточным качеством.
- Задача усложняется отсутствием разметки и невозможностью до-обучить систему распознавания текста на почерке автора сочинения.

Цель Разработка метода обнаружения почти-дубликатов в изображениях рукописного текста.

Приложение Использование системы для поиска заимствований в рукописных сочинениях школьников.

Постановка задачи

$\mathcal{D}_{\text{susp}} = \{d_{\text{susp}}^i\}_{i=1}^n$ — множество рукописных текстов для проверки.

$\mathcal{D} = \{d^j\}_{j=1}^m$ — множество текстовых документов коллекции.

Документ $d_{\text{susp}}^i \in D_{\text{susp}}$ имеет один источник заимствования $g : D_{\text{susp}} \rightarrow D$.

Требуется найти отображение $f : D_{\text{susp}} \rightarrow \{0, 1\}^D$, решая задачу:

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

где \mathcal{F} — множество рассматриваемых моделей,

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

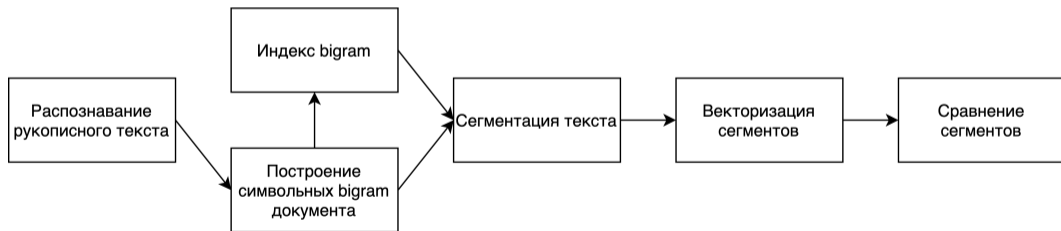
где $f(d_{\text{susp}}^i) @K$ — множество из K наибольших элементов коллекции для порядка, индуцированного $f(d_{\text{susp}}^i)$.

- Извлечем границы слов, содержащихся в сочинении **без их распознавания**.
- Получим признаки, инвариантные для различных почерков:
 - Нормализованная длина слова.
 - Нормализованная высота слова.
 - Наличие лигатур.
- Итоговый алгоритм поиска почти-дубликатов заключается в сопоставлении полученных последовательностей признаков, извлеченных из текста.

Распознавания слов:

- Бинаризация изображения и выделения строк.
- Сегментация слов на основе поиска компонент связности двухкомпонентной гауссовой смеси.
- Расстояние между последовательностями задается функцией Dynamic time warping.

- Распознавание рукописного текста (предобученная модель)¹.
- Построение символьных n-gram над распознанными символами.
- Поиск документов-кандидатов на основе полученных символьных n-gram.
- Векторизация текстовых сегментов и их попарное сопоставление (bag of n-gram)².



¹ https://github.com/ai-forever/htr_datasets

² Использовались доверительные n-gram, которые имеют высокое качество распознавания

Коллекция документов на основе выборки Тайга³:

- Объем коллекции: 76855 документ.
- Препроцессинг:
 - Размер текста не менее 20,000 символов.
 - Число слов на русском языке не менее 100.

Разнородные типы текстовых изображений:

- Тексты написаны на бланках и отсканированы⁴.
- Тексты написаны на произвольных бланках/листах и сфотографированы.

³ *Shavrina T., Shapovalova O.* To the methodology of corpus construction for machine learning: “taiga” syntax tree corpus and parser // Proceedings of the “Corpora”: 78-84, 2017.

⁴ *Bakhteev O., et al* Near-duplicate handwritten document detection without text recognition // Computational Linguistics and Intellectual Technologies, 2021.

Примеры изображений с текстом

Итоговое сочинение (изложение)
Бланк записи

Код школы: 25
Код региона: 25
ФИО: _____
Дата: _____

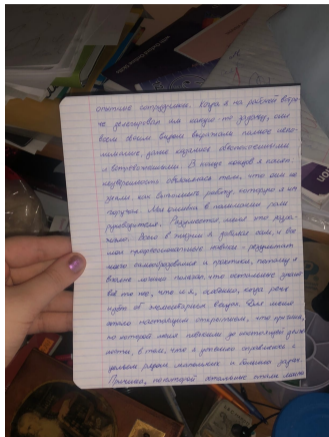
ПРИМЕЧАНИЕ Все бланки регистрируются и оцифровываются.

Напишите, на какие две задачи, которые не исполнили все о работе. Какой этап из оформления написание работы и какие требования. Сначала делается работа над текстом и только потом, пишется текст на тему работы. Сначала, делается в основном текст, только текст, который можно сделать. Потом можно сделать требования, которые к работе, и оформление работы. Не надо писать на работе. Каким образом работа оформляется требования, и требования, на работу, оформляется работа.

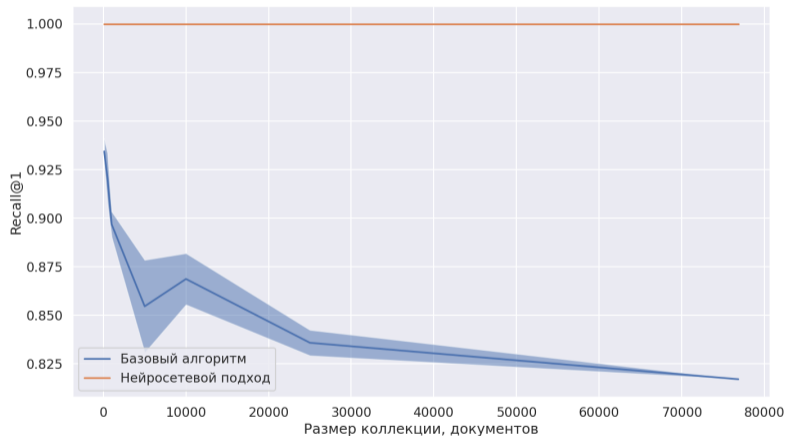
В рассказе А.С. Пушкина «Три года в ссылке» писатель А.А. Дельвиг описывает деятельность журналиста в ссылке, в которой была. Все это делается, при котором нужно соблюдать требования к работе. «Два года в ссылке», — пишет Дельвиг о деятельности журналиста в ссылке. «Два года в ссылке», — пишет Дельвиг о деятельности журналиста в ссылке. «Два года в ссылке», — пишет Дельвиг о деятельности журналиста в ссылке. «Два года в ссылке», — пишет Дельвиг о деятельности журналиста в ссылке.

В рассказе А.С. Пушкина «Три года в ссылке» писатель А.А. Дельвиг описывает деятельность журналиста в ссылке, в которой была. Все это делается, при котором нужно соблюдать требования к работе. «Два года в ссылке», — пишет Дельвиг о деятельности журналиста в ссылке. «Два года в ссылке», — пишет Дельвиг о деятельности журналиста в ссылке. «Два года в ссылке», — пишет Дельвиг о деятельности журналиста в ссылке. «Два года в ссылке», — пишет Дельвиг о деятельности журналиста в ссылке.

Для удобства места для записи оформлены в соответствии с требованиями.



Слева: сканы бланков. Справа: произвольные фотографии с текстом.



Нейросетевой подход имеет значительное преимущество над базовым подходом.

Модель	Бланки государственного образца	Произвольные бланки
Базовая модель	81,6	18,2
Нейросетевой подход	100,0	61,7

При изменении качества входных данных, качество **базовой модели** значительно ухудшается.

Нейросетевой подход имеет все еще высокое качество.

- Проведено тестирование базового метода поиска почти дубликатов рукописных текстов на разнородных изображениях.
- Предложен метод поиска почти дубликатов на основе нейросетевого подхода.
- Анализ предложенного нейросетевого подхода показал его устойчивость к изменению формата данных.
- В дальнейшем планируется расширение выборки разнородных снимков рукописных текстов для увеличения качества работы нейросетевой модели.