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

Задача поиска почти-дубликатов рукописных текстов

Проблема

- Во многих отраслях все еще присутсвуют 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

$\mathfrak{D}_{\text{susp}} = \{d^i_{\text{susp}}\}_{i=1}^{n}$ — множество рукописных текстов для проверки.
$\mathfrak{D} = \{d^j\}_{j=1}^{m}$ — множество текстовых документов коллекции.

Документ $d^i_{\text{susp}} \in D_{\text{susp}}$ имеет один источник заимствования $g : D_{\text{susp}} \to D$.

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

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

где $\mathcal{F}$ — множество рассматриваемых моделей,

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

где $f\big(d^i_{susp}\big)@K$ — множество из K наибольших элементов коллекции для порядка, индуцированного $f\big(d^i_{susp}\big)$.

Решение на основе характеристик текста

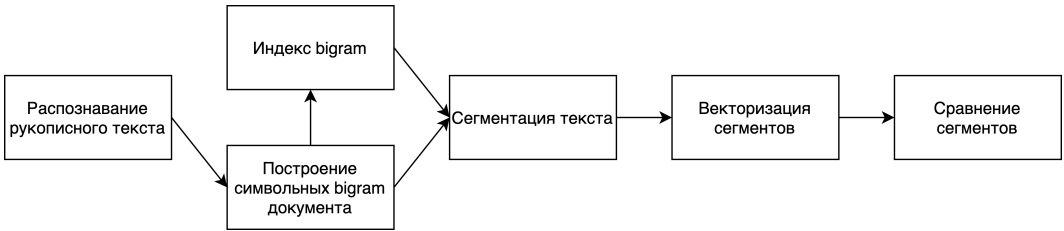
- Извлечем границы слов, содержащихся в сочинении **без их распознавания**.
- Получим признаки, инвариантные для различных почерков:
  - Нормализованная длина слова.
  - Нормализованная высота слова.
  - Наличие лигатур.
- Итоговый алгоритм поиска почти-дубликатов заключается в сопоставлении полученных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

Нейросетевой подход

- 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).



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

Вычислительный эксперимент

Коллекция документов на основе выборки Тайга[2]:

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

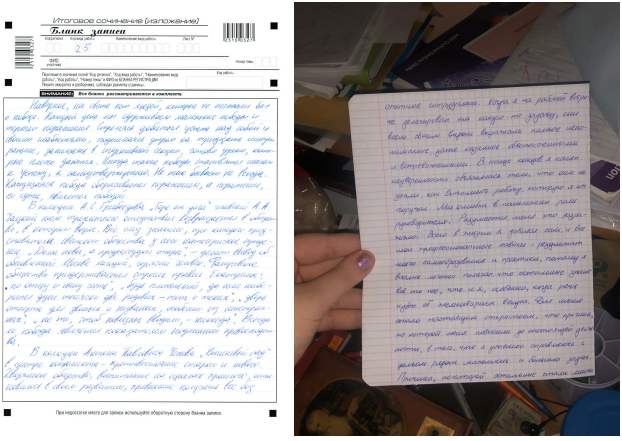
Разнородные типы текстовых изображений:

- Тексты написаны на бланках и отсканированы[3].
- Тексты написаны на произвольных бланках/листах и сфотографированы.

[2] *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.

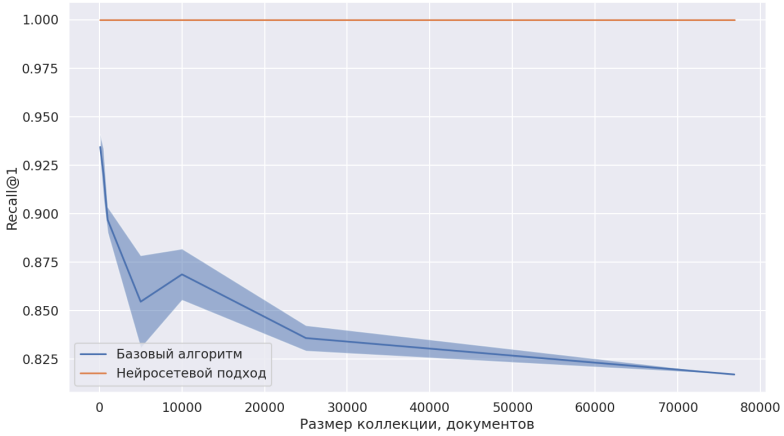
[3] *Bakhteev O., et all* Near-duplicate handwritten document detection without text recognition // Computational Linguistics and Intellectual Technologies, 2021.

Примеры изображений с текстом



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

Качество на отсканированных бланках



Нейросетевой подход имеет значительное преимущество над базовым подходом.

Качество на произвольных бланках и страницах

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.

Заключение

- 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.