

РАЗРАБОТКА МОДУЛЯ ДЛЯ ПРЕДОБРАБОТКИ ДОКУМЕНТОВ ОБНАРУЖЕНИЯ ЗАИМСТВОВАНИЙ В ТЕКСТЕ

Крез К. С., Ефименко Д. Д., Войнилович Н. Ю.

Кафедра проектирования информационно-компьютерных систем,
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: k.krez@bsuir.by

*В статье рассматриваются основные этапы предварительной обработки *.pdf и *.docx-документов для обнаружения заимствований в тексте и принципы их работы. Выявлены ключевые проблемы предварительной обработки *.pdf и *.docx-документов. Предложены пути их решения в том числе использование специализированных библиотек для языка Python.*

ВВЕДЕНИЕ

В условиях роста объемов цифровой информации и широкого распространения автоматизированных систем анализа текстов особую значимость приобретает задача эффективной предварительной обработки данных. Одним из ключевых этапов данного процесса является очистка текста, направленная на удаление «шумовых» и «нерелевантных» элементов, препятствующих корректному проведению лингвистического и семантического анализа [1].

Особую сложность представляет обработка файлов в форматах *.pdf и *.docx, поскольку они включают не только текстовые данные, но и графические материалы, метаданные и элементы верстки. В *.pdf-документах текст, как правило, представлен в виде множества независимых блоков, разделенных линиями и координатными сетками, что делает затруднительным его прямое использование в алгоритме. В документах *.docx значительная часть информации содержится в технических метках и стилях, не имеющих смысловой нагрузки, но влияющих на корректность извлечения контента.

В связи с этим актуальной задачей является разработка модуля, обеспечивающих автоматическую очистку текстов от технических и структурных артефактов, с сохранением только релевантной смысловой информации.

I. РЕАЛИЗАЦИЯ ПРОЦЕССА ОЧИСТКИ

Процесс очистки текстовых данных в разработанной системе реализован в модуле text_cleaning.py [2], который представляет собой совокупность последовательно выполняемых процедур, направленных на удаление «шумовых, графических, технических и структурных» элементов документа [3]. Реализация данного модуля основана на принципах модульности, расширяемости и независимости от конкретного формата исходного файла. Это позволяет применять его как к документам формата *.pdf, так и к *.docx, обеспечивая единый подход к формированию очи-

щенного текста. Этапы очистки текстового документа:

1. Удаление ссылок на графические элементы направлен на исключение всех упоминаний графических объектов – рисунков, схем, диаграмм и таблиц, – которые не несут содержательной информации.

Для удаления подобных ссылок используются регулярные выражения, позволяющие находить и удалять подписи и текстовые отсылки к иллюстрациям. Наиболее часто применяемые паттерны:

- $r'(Рисунок|Таблица|рисунок|таблица) \backslash s^* \backslash d+ ([?]\backslash d+)^* \backslash s^* [?]\backslash s^* .*$ – удаление подписей к рисункам и таблицам;
- $r'\backslash b(на\ рисунке|На\ рисунке|На\ рисунках) \backslash b.*?[?]\backslash n.]^*$ – удаление ссылок на графические элементы внутри текста.

2. Удаление технических элементов документа часто присутствуют в отчетах, учебных и нормативных документах. Они не несут смысловой информации, однако могут существенно влиять на корректность токенизации и разметки текста. Для устранения подобных артефактов используются следующие паттерн:

- удаление номеров страниц: $r'\backslash n \backslash s^* \backslash d+ \backslash s^* \backslash n$;
- удаление стандартных заголовков (например, шапок государственных документов): $r'Министерство образования Республики Беларусь.*?Минск \backslash s^* \backslash d\{4\}$;
- удаление служебных обозначений и маркировок: $r'IP \backslash s+ [A-Я](\backslash s+ [-])?'$.

3. Удаление структурных элементов документа связанных с организационной структурой документа, таких как оглавления (содержание), списки использованных источников, приложения, ведомость и т.д. Эти части не несут содержательной значимости и лишь дублируют информацию, уже содержащуюся в структуре текста.

Удаление осуществляется по ключевым словам (например, «СОДЕРЖАНИЕ», «СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ», «ПРИЛОЖЕНИЕ», «ВЕДОМОСТЬ») и по регулярным выражениям, описывающим строки, содержащие только последовательности точек.

4. Нормализация пробелов и переносов строк, когда после удаления структурных и технических элементов выполняется нормализация текста. Цель данной этапа – обеспечение единообразного форматирования для последующих этапов анализа.

Множественные пробелы, табуляции и повторяющиеся переносы строк заменяются на одиночные пробелы с помощью регулярного выражения $r'\backslash s^+$. В результате формируется ровный, непрерывный текст без лишних разрывов и неформатированных фрагментов.

5. Обработка заголовков и структурирование текста направлены на обеспечение логической целостности документа посредством корректной идентификации разделов и глав. Функция `is_chapter_heading()` анализирует каждый текстовый блок на соответствие определённым паттернам, отражающим структуру документа:

- $r'\backslash d+\backslash s+[A-Я\backslash s]^+\$'$ – нумерованные заголовки, записанные прописными буквами;
- $r'\backslash d+\backslash \. \backslash d+\backslash s+[A-Я][a-я\backslash s]^+$ – иерархические заголовки;
- $r'\backslash [A-Я\backslash s]^+\$'$ – ненумерованные заголовки, оформленные заглавными буквами.

На основе найденных заголовков выполняется разделение текста по разделам, где каждая глава представлена заголовком и соответствующим ей очищенным содержанием.

В отличие от *.pdf и *.docx, где очистка выполняется после извлечения содержимого, для файлов формата *.docx очистка интегрирована непосредственно в процесс извлечения. Система извлекает только текстовые параграфы, игнорируя технические метки и элементы форматирования, не влияющие на смысловую структуру. Это способствует значительному снижению объёма

постобработки и повышает качество итогового текста.

II. ПРАКТИЧЕСКАЯ ЧАСТЬ

Практическая часть исследования направлена на комплексную оценку производительности разработанного модуля текстовой очистки `text_cleaning.py`. Для обеспечения репрезентативности результатов была сформирована тестовая выборка, включающая: тип документа (курсовая и дипломная работа), формат документа (*.pdf и *.docx), кол-во страниц.

ЗАКЛЮЧЕНИЕ

Представленный модуль очистки *.pdf и *.docx документа использует минимальный набор этапов таких как: удаление ссылок на графические элементы, удаление технических элементов документа, удаление структурных элементов документа, нормализация пробелов и переносов строк, обработка заголовков и структурирование текста. В ходе работы была рассмотрена реализация данного процесса в рамках модуля `text_cleaning.py`.

Практическая часть показала, что: статистически значимые показатели производительности модуля со средним временем обработки $3,31 \pm 0,02$ мс/страницу; эмпирически подтверждена линейная зависимость временных характеристик обработки от объема документа.

Таким образом, предложенное решение представляет законченный модуль предобработки текстовых данных, обладающую теоретической обоснованностью и практической эффективностью для интеграции в программное средство для проверки академической добросовестности.

1. Разработка программного модуля распознавания документов на основе машинного обучения [Электронный ресурс]. – Режим доступа: <https://apni.ru/article/11072-razrabotka-programmnogo-modulya-raspoznaniya-dokumentov-na-osnove-mashinnogo-obucheniya>. – Дата доступа: 14.09.2023.
2. Textclean: Text cleaning tools [Electronic resource]. – Mode of access: <https://github.com/trinker/textclean>. – Date of access: 14.09.2023.
3. Хобсон Лейн, Ханнес Хапке, Коул Ховард. Обработка естественного языка в действии / Л. Хобсон, Х. Хапке, К. Ховард. – СПб.: Питер, 2020. – С. 68–140.

Таблица 1 – Результаты очистки текста в документах

Тип документа	Кол-во стр.	Измерение, мс.						Сред. знач., мс.
		1	2	3	4	5	6	
Курс. работа 1 (*.pdf)	67	374,94	374,83	374,93	374,83	374,76	374,85	374,87
Курс. работа 2 (*.docx)	67	379,23	379,12	379,24	379,13	379,21	379,23	379,19
Диплом. работа 1 (*.pdf)	89	387,34	387,31	387,29	387,34	387,31	387,31	387,31
Диплом. работа 2 (*.docx)	89	387,59	387,61	387,58	387,59	387,59	387,61	387,60