Программные инструменты поиска близких фрагментов в коллекции учебных документов

С.М. Бормотова[0009-0001-6400-1980], О.А. Невзорова[0000-0001-8116-9446]

КФУ /Институт вычислительной математики и информационных технологий

Аннотация. В статье рассматриваются методы автоматического разделения электронных документов в формате docx и pdf на логические фрагменты (разделы и главы) с последующим семантическим анализом с помощью модели Sentence-BERT. Первый метод выявляет и анализирует изменения шрифта и стилей текста, таким образом обнаруживая начало главы или раздела. Второй – конвертирует документ pdf в docx, находит главы со служебными словами по типу «Глава», «Содержание» и др., найденные главы выводит в графический интерфейс с последующей ручной настройкой, что эффективно для файлов, которые имеют нестандартную разметку. Третий метод использует встроенные закладки в документе, что обеспечивает максимальную точность для документов с предопределенной структурой. Для семантического анализа фрагментов используется модель Sentence-BERT, которая выявляет смысловые связи между главами и разделами. Приведены результаты тестирования на 74 документах, которые показывают преимущества и недостатки каждого метода. Разработанные инструменты автоматизируют процесс обработки документов, позволяя быстро делить коллекцию документов на смысловые фрагменты для поиска схожих фрагментов в ней, а также провести анализ найденных близких по содержанию пар с помощью графического интерфейса.

Ключевые слова: обработка документов, семантический анализ, разделение на главы, PDF, Sentence-BERT.

Software tools for searching for similar fragments in a collection of educational documents

S.M. Bormotova, O.A. Nevzorova

KFU/Institute of Computational Mathematics and Information Technologies

Abstract. The article examines methods for automatic segmentation of electronic documents in docx and pdf formats into logical fragments (sections and chapters) with subsequent semantic analysis using the Sentence-BERT

model. The first method detects and analyzes changes in text fonts and styles, thereby identifying chapters or sections. The second method converts pdf documents to docx, locates chapters using structural markers such as "Chapter," "Contents," etc., and displays the identified chapters in a graphical interface for subsequent manual adjustment, which is particularly effective for files with non-standard layouts. The third method utilizes built-in document bookmarks, ensuring maximum accuracy for documents with predefined structures. For semantic analysis of fragments, the Sentence-BERT model is employed to identify meaningful relationships between chapters and sections. Testing results on 74 documents are presented, demonstrating the advantages and limitations of each method. The developed tools automate the document processing workflow, enabling rapid segmentation of document collections into meaningful fragments for identifying similar content, as well as analysis of found content-related pairs through a graphical interface.

Keywords: document processing, semantic analysis, text segmentation, PDF, Sentence-BERT

1. Введение

Учебники и учебные пособия являются основным источником знаний учащихся и одним из важнейших средств обучения. Учебники отличаются способом изложения учебного материала в зависимости от возраста и подготовки учащихся Усвоение учебного материала – это важнейшая составляющая процесса обучения, для успешного усвоения разрабатываются различные методики. Персонализированный подход к обучению включает этапы, связанные с оценкой уровня подготовки обучаемого и предъявления ему учебного материала в соответствии с уровнем подготовки. Таким образом, актуальной является задача поиска и оценивания в учебной тематической коллекции документов фрагментов, близких по содержанию, относящихся к одной учебной теме. Сравнительная оценка найденных фрагментов по заданным критериям позволит выбирать для изучения тематические фрагменты, соответствующие уровню подготовки учащихся. В статье рассматривается только задача поиска близких фрагментов в коллекции учебных материалов, включая основные методы обработки документов в формате pdf и построение программных инструментов для решения указанной задачи поиска.

Решение для задачи поиска близких фрагментов в коллекции учебных материалов строится в два этапа. На первом этапе выполняется автоматическое разделение документа в формате pdf на фрагменты с помощью разработанных методов, на втором этапе выполняется сравнение фрагментов на близость содержания на основе модели Sentence-BERT.

1. Методы разделения pdf-документов на фрагменты

Автоматическое разделение pdf-документов на части решается в настоящее время различными доступными программными инструментами, однако все эти решения требуют участия пользователя на этапе настройки соответствующей функции. Решение, разработанное в данной статье, не требует участия пользователя и разделяет документ на части по заголовкам разных уровней.

Для решения поставленной задачи были разработаны три метода. Для обработки pdf-документа использована библиотека PyMuPDF, для документа в формате docx – библиотека python-docx. Основной (первый) метод анализирует изменения шрифта в тексте и находит отклонения стиля и размера шрифта от основного шрифта.

Для файла формата docx распознаются стили и размеры шрифта заголовков первого и второго уровня (Heading 1 и Heading 2), а также выделяются характерные слова для логического деления, такие как «Глава», «Введение» и другие. Далее эти признаки идентифицируются как новый заголовок, по которому в дальнейшем документ разделяется на фрагменты. Второй метод является модификацией первого метода и предназначен для разделения файла в формате pdf, для которого вначале выполняется предобработка, связанная с удалением изображений, а затем его конвертация в формат docx.

Ключевые этапы обработки документа формата docx:

- Загрузка и валидация - проверка файлов на целостность и соответствие формату;
- Предобработка – очистка от нестандартных символов и медиа-контента;
- Анализ стилевых характеристик – распознавание размера шрифта и поиск ключевых слов («Глава», «Содержание» и др.);
- Разделение на логические фрагменты по выставленным ключевым точкам на предыдущем этапе;
- Постобработка – удаление пустых страниц, проверка читаемости.

Преимуществом разработанных методов является возможность установить ключевые точки разбиения, даже если в документе нет явных ссылок на логическое деление, т.е. отсутствуют характерные ключевые слова. Недостаток методов связан с низкой точностью (менее 70%) и вычислительными затратами.

Третий метод разделения документа распознает встроенные закладки с помощью готового инструмента PDFSam, который извлекает оглавление из метаданных документа [1].

Преимуществами этого метода являются высокая точность, так как документ разделяется по заранее созданным автором меткам и высокая скорость обработки. Существенным недостатком является зависимость от качества разметки книги и наличия закладок.

2. Эксперименты

Исследование проводилось на коллекции документов из 74 учебников по информатике и программированию. С помощью первого алгоритма обработана 31 книга, с помощью второго, расширяющего функционал первого – 49 книг, с помощью третьего – 7 книг.

Ниже представлены таблицы сравнения методов на трех книгах с закладками.

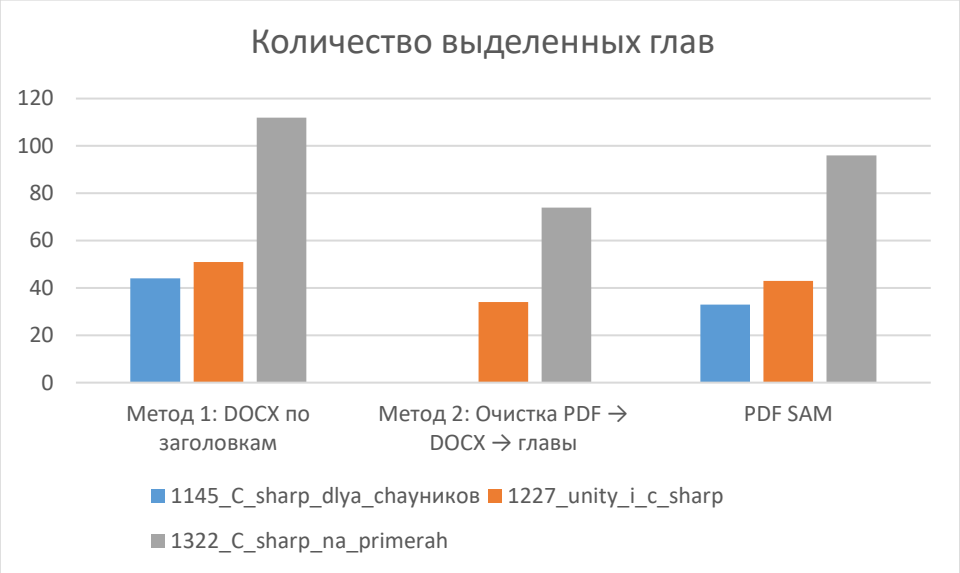

Таблица 1. Количество выделенных глав каждым методом

Время обработки			
Метод обработки	1145_C_sharp_ dlya_chайников	1227_u nity_i_c_sharp	1322_C_sha rp_na_primerah
Метод 1: DOCX по заголовкам	2,03	2,20	1,89
Метод 2: Очистка PDF → DOCX → главы	228,00	192,00	231,00
PDF SAM	5,42	4,17	3,74

Таблица 2. Время (в секундах) обработки книг методами

Проведенное исследование показало, что выбор оптимального метода разделения документов на главы зависит от типа исходных файлов и требований к скорости обработки. Метод на основе анализа шрифтов (Метод 1) демонстрирует наилучшее сочетание скорости (1-3 секунды на документ) и приемлемой точности, что делает его предпочтительным для массовой обработки хорошо структурированных документов. Библиотека PDFSam обеспечивает максимальную точность при работе с файлами, содержащими закладки, однако применим лишь к 66% тестовой выборки.

Наиболее универсальным, но ресурсоемким (до 4 минут на файл) оказался метод 2, сочетающий конвертацию формата pdf в docx с ручной корректировкой. Его ключевое преимущество – способность обрабатывать документы со сложной или поврежденной структурой, недоступные для других методов. Для комплексного решения рекомендуется комбинированный подход: первоначальная обработка с помощью PDFSam (если имеются закладки), в случае отсутствия закладок использовать метод 1, а для проблемных файлов – метод 2 с последующей верификацией результатов.

2. Семантический анализ текстовых фрагментов с использованием модели Sentence-BERT

Для выявления смысловых связей между фрагментами различных учебников использовался метод семантического сравнения на основе модели Sentence-BERT [2]. Данный метод позволяет находить близкие по содержанию фрагменты даже при различиях в формулировках и стиле изложения.

Процесс анализа включает несколько ключевых этапов. Сначала происходит извлечение текста из подготовленных фрагментов документов. Затем с помощью предобученной multilingual модели paraphrase-MiniLM-L12-v2 каждый текст преобразуется в векторное представление (эмбеддинг) [3]. Особенностью выбранной модели является ее способность сохранять смысловые связи между предложениями на разных языках, что важно для анализа учебной литературы.

Для количественной оценки схожести фрагментов используется метрика косинусного сходства между векторами. Установленный порог в 0.6 позволяет отфильтровывать случайные совпадения, оставляя только значимые смысловые связи. Анализ проводится в четырех направлениях: сравнение разделов внутри одной книги, между разными учебниками, а также сопоставление глав и подразделов различных уровней.

Результаты визуализируются с помощью специально разработанного графического интерфейса (рис. 1), который позволяет экспертам просматривать пары схожих фрагментов и оценивать качество автоматического анализа. Гистограммы распределения значений сходства

(рис. 2, рис. 3) демонстрируют высокую согласованность структуры внутри отдельных учебников (пики в области 0.8-1.0) и умеренное, но статистически значимое сходство между разными изданиями (диапазон 0.6-0.75).

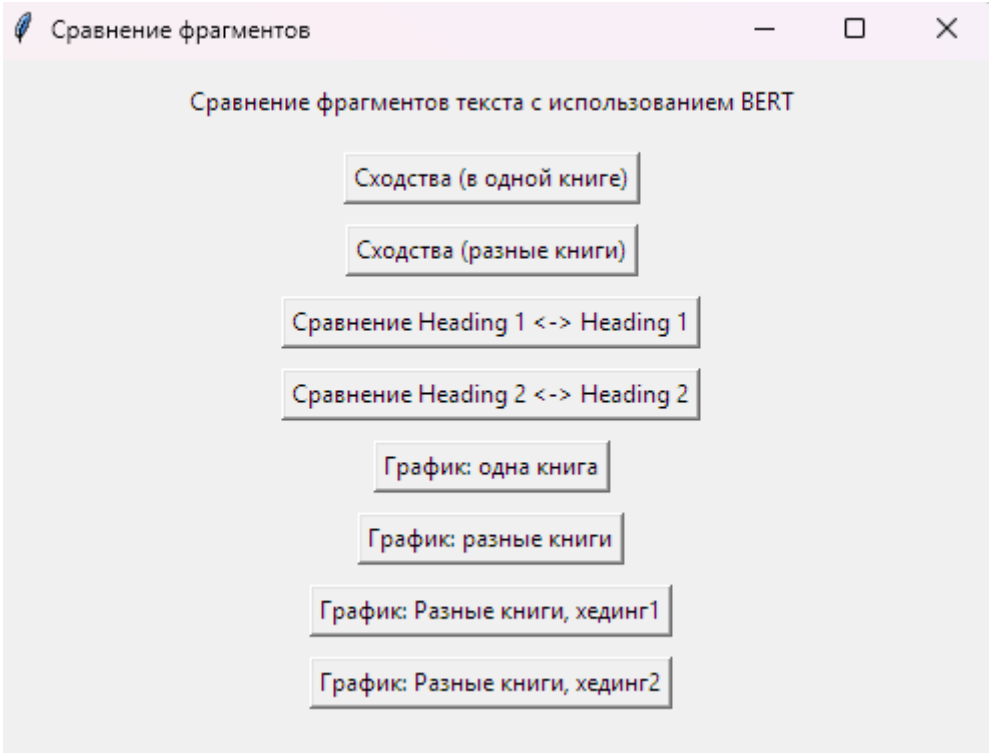

Рис. 1. Интерфейс программы для сравнения содержания фрагментов

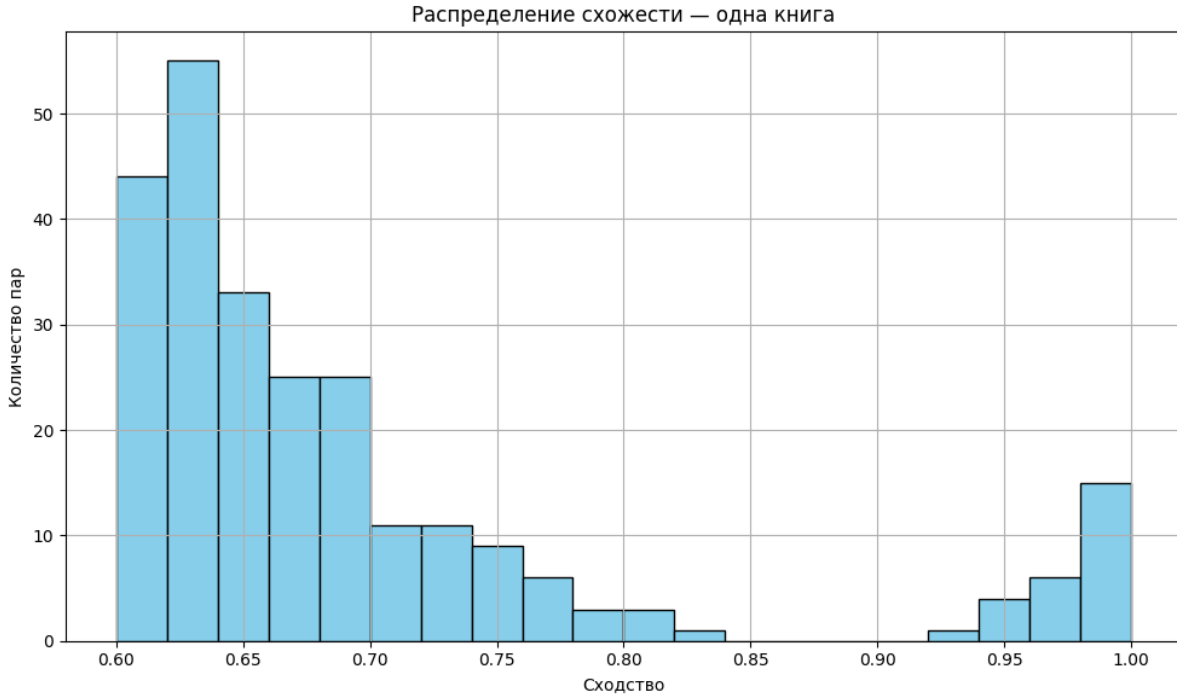

Рис. 2. Гистограмма распределения близости фрагментов в одной книге

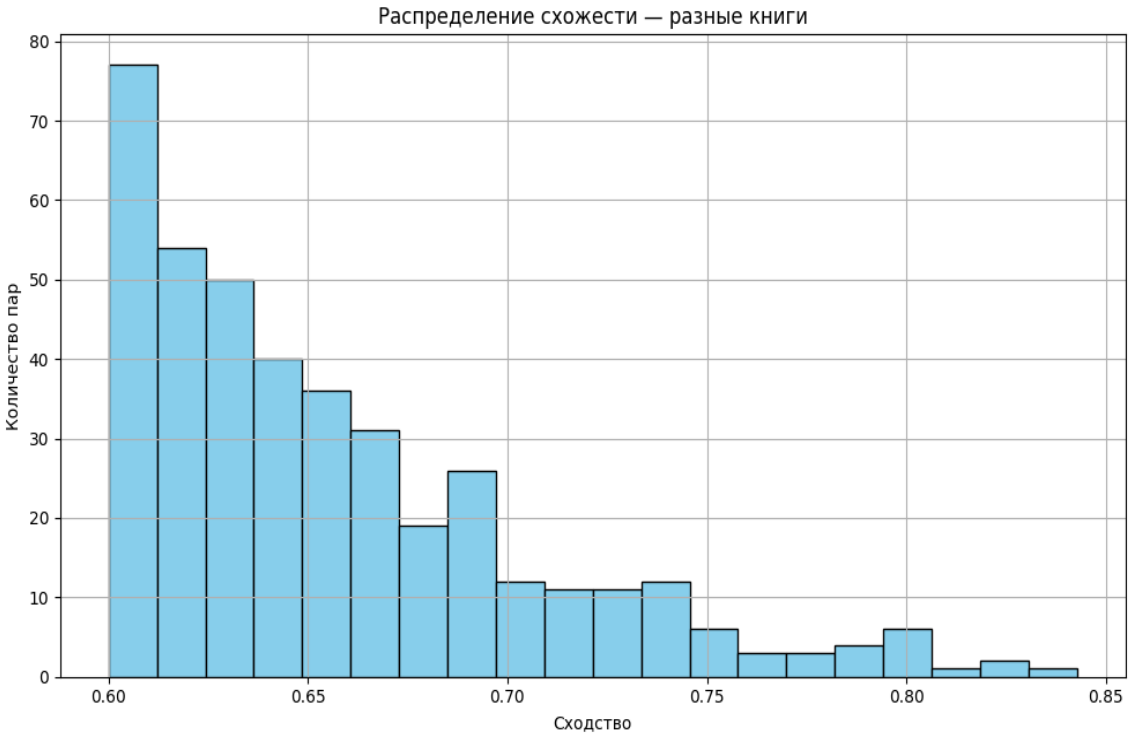

Рис. 3. Гистограмма распределения близости фрагментов в разных книгах.

Анализ позволил выявить ряд закономерностей. Так, на рисунке 2 внутри одной книги выявлены высокие показатели близости выделенных фрагментов (глав), что указывает на согласованность структуры внутри одного учебника, так что заголовки второго уровня логично дополняют заголовки первого уровня, раскрывая их содержание. При анализе рисунка 2 можно также увидеть, что оценки близости пар фрагментов попадают в диапазон 0.6-0.8, что означает частичное совпадение тем в разных учебниках. Это может означать, что авторы используют похожие логические структуры или некоторые темы действительно распространены и появляются в похожей форме в разных книгах.

Заключение

В статье рассмотрены методы автоматического разделения электронных документов в формате docx и pdf на логические фрагменты (разделы и главы) с последующим семантическим анализом с помощью модели Sentence-BERT. Разработан программный инструмент, который может анализировать и сопоставлять текстовые данные из разных учебных документов. Применение семантического анализа существенно расширяет возможности разработанного инструментария, позволяя не только разделять документы на структурные элементы, но и выявлять содержательные связи между материалами из разных источников. Это особенно ценно для образовательных платформ, где важно находить и сопоставлять аналогичные темы в различных учебных пособиях.

Литература

1. Программы для обработки PDF: основные функции. [Электронный ресурс] – URL:https://store.softline.ru/blog/design/programmy-dlya-obrabotki-pdf/ (дата обращения 20.07.2025).
2. PDF Split and Merge (PDFSAM): открытый инструмент для обработки PDF-документов. [Электронный ресурс] – URL: https://github.com/torakiki/pdfsam (дата обращения 20.07.2025).
3. Sentence Transformers Documentation. [Электронный ресурс] – URL: https://sbert.net/(дата обращения 20.07.2025).
4. Shundeev A., Balakhnichev S., Zaslavskii D., Pekhterev S. Document classification using word embeddings // Proceedings of the 6th International Conference Actual Problems of System and Software Engineering Moscow, Russia, 2019. Pp. 377-388. – https://ceur-ws.org/Vol-2514/

References

1. PDF Processing Software: Basic Functions. – URL:https://store.softline.ru/blog/design/programmy-dlya-obrabotki-pdf/ (accessed on July 20, 2025).
2. PDF Split and Merge (PDFSAM): An open source tool for processing PDF documents. – URL: https://github.com/torakiki/pdfsam (accessed on July 20, 2025).
3. Sentence Transformers Documentation. – URL: https://sbert.net/ accessed on July 20, 2025).
4. Shundeev A., Balakhnichev S., Zaslavskii D., Pekhterev S. Document classification using word embeddings // Proceedings of the 6th International Conference Actual Problems of System and Software Engineering Moscow, Russia, 2019. Pp. 377-388. – https://ceur-ws.org/Vol-2514/