Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина)

Разработка системы автоматизированной проверки наиболее частых ошибок в научных текстах

Выполнил: Блеес Эдуард Игоревич, гр. 3304
Руководитель: Заславский Марк Маркович, ассистент

Актуальность

Процесс проверки статей изданиями в текущем виде:

- Долгая переписка с рецензентом и редакторами;
- Повторные отправки после малейших исправлений.

Существует курс по написанию научных статей на Stepik, для которого необходима частичная автоматизация проверки статей.

Цель и задачи

Цель: Разработать программу для проверки статьи на соответствие научному стилю и поиску наиболее частых ошибок в ней.

Задачи:

- Исследовать возможность автоматизации проверки научных статей на соответствие научному стилю;
- Построить математическую модель проверки статьи;
- Провести экспериментальное исследование для определения допустимых значений критериев;
- Реализовать программный прототип решения.

Исследование. Научный стиль. Применимость решения.

Научный стиль - наиболее строгий стиль речи в русском языке. Собственно-научный подстиль — академическое изложение, адресованное специалистам. Морфологические ограничения.

Разработка модели направлена на проверку **необходимых, но не достаточных** условий соответствия текста научному стилю. Автоматическая оценка **научной ценности** статьи – не решенная на данный момент задача. Над подобной задачей работают Google и другие компании.

Исследование возможности автоматизации проверки статей на соответствие научному стилю

Проверка качества текста или соответствие информационному стилю. SEO-анализ.

Вводимые термины:

- **Плотность текста** – это показатель повторений в текстовом документе ключевых слов и фраз. Используется и другое название – «тошнота» текста.
- **Стоп-слова** – это слова в тексте, которые не несут смысловой нагрузки (предлоги, союзы, частицы и т.п.)
- **Вода** - процентное соотношение стоп-слов и общего количества слов в тексте
- Эмпирическая закономерность распределения частоты слов естественного языка - **Закон Ципфа**

Исследование возможности автоматизации проверки статей на соответствие научному стилю

Обзор аналогов

Аналог	Многокритериальная проверка	Нет ограничения на длину текста	Проверка стиля	Возможность загрузки файлов для проверки
1y.ru	-	+	-	-
text.ru	+	-	-	-
content monster .ru	+	+	-	-
glvrd.ru	+	+	+	-

Экспериментальное исследование

Выборка из 2500 статей опубликованных в источниках ВАК или РИНЦ.

Проверяемая гипотеза:
Качество научной статьи влияет на значения определенных числовых критериев, а также полученная выборка значений критериев соответствует нормальному распределению.

Результаты экспериментального исследования

Числовые критерии:

- Тошнота/Плотность текста – α;
- Уровень воды в тексте – β;
- Значение отклонения текста статьи от идеальной кривой по Ципфу – λ.

Экспериментально установленные интервалы:

Критерий	Интервал
α	[6, 14]
β	[14, 20]
λ	[5.5, 9.5]

Математическая модель проверки статьи

В результате исследования было выделено:

- 3 рассчитываемых числовых критерия;
- 5 типов проверяемых стилистических ошибок;
- 6 типов проверяемых структурных ошибок.

Оценка статьи:

$$К = В - Ф$$

Где К – оценка статьи, В – базовое значение К, Ф – штраф.

Математическая модель проверки статьи

$$\mathrm{E}(\alpha) = \begin{cases} 1, \alpha \in [6;14] \\ 0, \alpha \notin [6;14] \end{cases}$$

$$\mathrm{B} = C_1 \times \mathrm{E}(\alpha) + C_2 \times \mathrm{E}(\beta) + C_3 \times \mathrm{E}(\lambda)$$

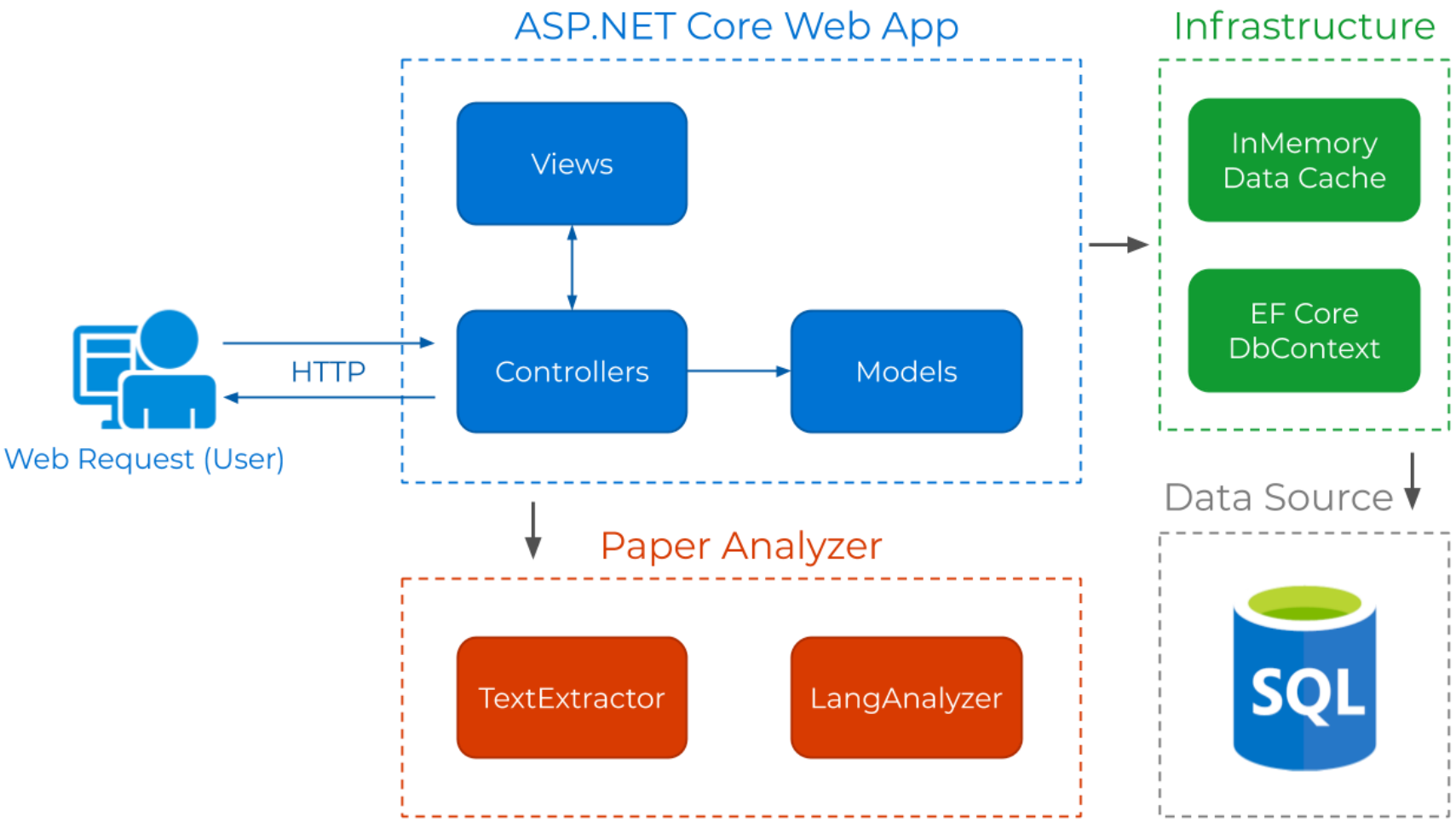
$$\Phi = C_4 \times N_1 + C_5 \times N_2$$

Где E – попадание критерия в установленный промежуток, N_1 – количество структурных ошибок, N_2 – количество стилистических ошибок, C_1, C_2, C_3, C_4, C_5 – коэффициенты.

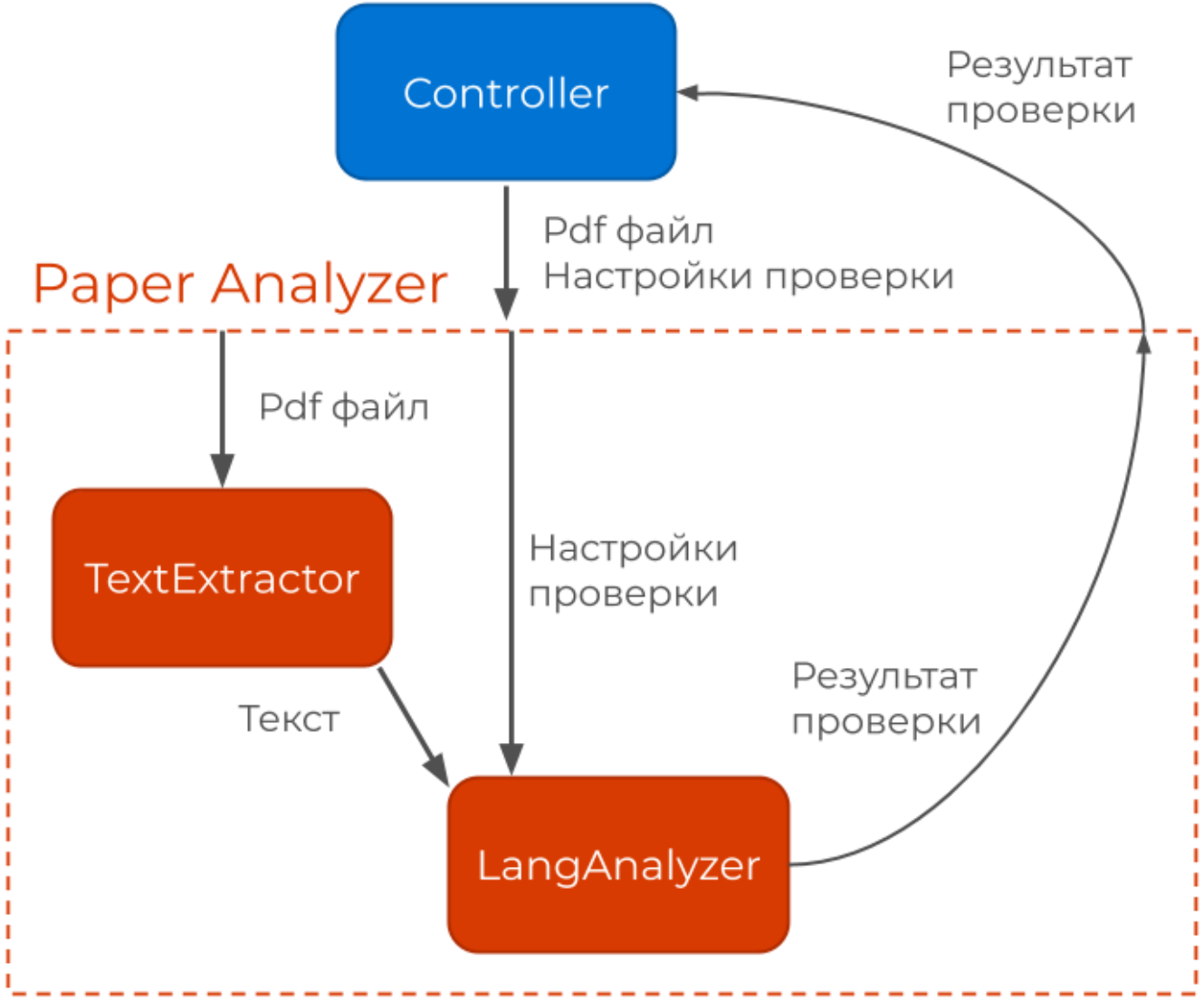
Разработанное решение

- Платформа .NET Core;
- Фреймворк ASP.NET Core MVC;
- Web API (REST) Controllers;
- SQL Server + ORM EF Core (реализован паттерн Repository, легко поддержать другую, в том числе не реляционную БД);
- Развернутое решение использует в среднем 450 Мбайт оперативной памяти, зафиксированная пиковая нагрузка не превышала 524 Мбайт;
- Среднее время обработки статьи на превышает 1 - 2 секунды.

Разработанное решение. Архитектура



Разработанное решение. Алгоритм обработки



Заключение

- Было проведено исследование возможности автоматизации проверки научных статей, была построена и настроена математическая модель проверки статьи;
- Было разработано решение в виде веб-сервиса.

Апробация работы

- Блеес Э.И., Заславский М.М. Автоматизация процесса проверки текста на соответствие научному стилю // Современные технологии в теории и практике программирования: материалы научно-практической конференции студентов, аспирантов и молодых ученых - 2018. - С. 118-121;
- Блеес Э.И., Заславский М.М. Исследование критериев соответствия текста научному стилю // Научно-технический вестник информационных технологий, механики и оптики. 2019. Т. 19. № 2. С. 299–305. doi: 10.17586/2226-1494-2019-19-2-299-305;
- Репозиторий проекта https://github.com/EduardBlees/Master-s-thesis.

Дополнительный слайд №1. Дальнейшее развитие

Развертывание с помощью docker и использование на кафедре для проверки статей из курса на Stepik

Дополнительный слайд №2. Почему не использовалось машинное обучение

Отсутствие корпуса для обучения.

Причина – политика изданий. Не был найден источник научных статей не прошедших рецензирование.

Дополнительный слайд №3. Поддержка английского языка

Морфологический анализ английского готов.
Необходимо:

- Провести исследование научного стиля английского языка;
- Выделить морфологические особенности научного стиля английского языка;
- Получить выборку опубликованных статей на английском языке и провести анализ числовых критериев для настройки модели проверки.

Дополнительный слайд №4. Примеры анализа текстов других жанров

Текст	α	$\alpha \in [6;14]$	β	$\beta \in [14,20]$	λ	$\lambda \in [5.5,\ 9.5]$
Псевдонаучная статья «Корчеватель»	10.38	Да	18.50	Да	6.84	Да
Интернет-статья «Моё разочарование в софте»	3.66	Нет	31.68	Нет	5.35	Нет
«Капитал» Карла Маркса	5.84	Нет	28.94	Нет	138.22	Нет
«Идиот» Фёдора Достоевского	6.65	Да	45.65	Нет	53.12	Нет

Дополнительный слайд №5. Проверяемые ошибки

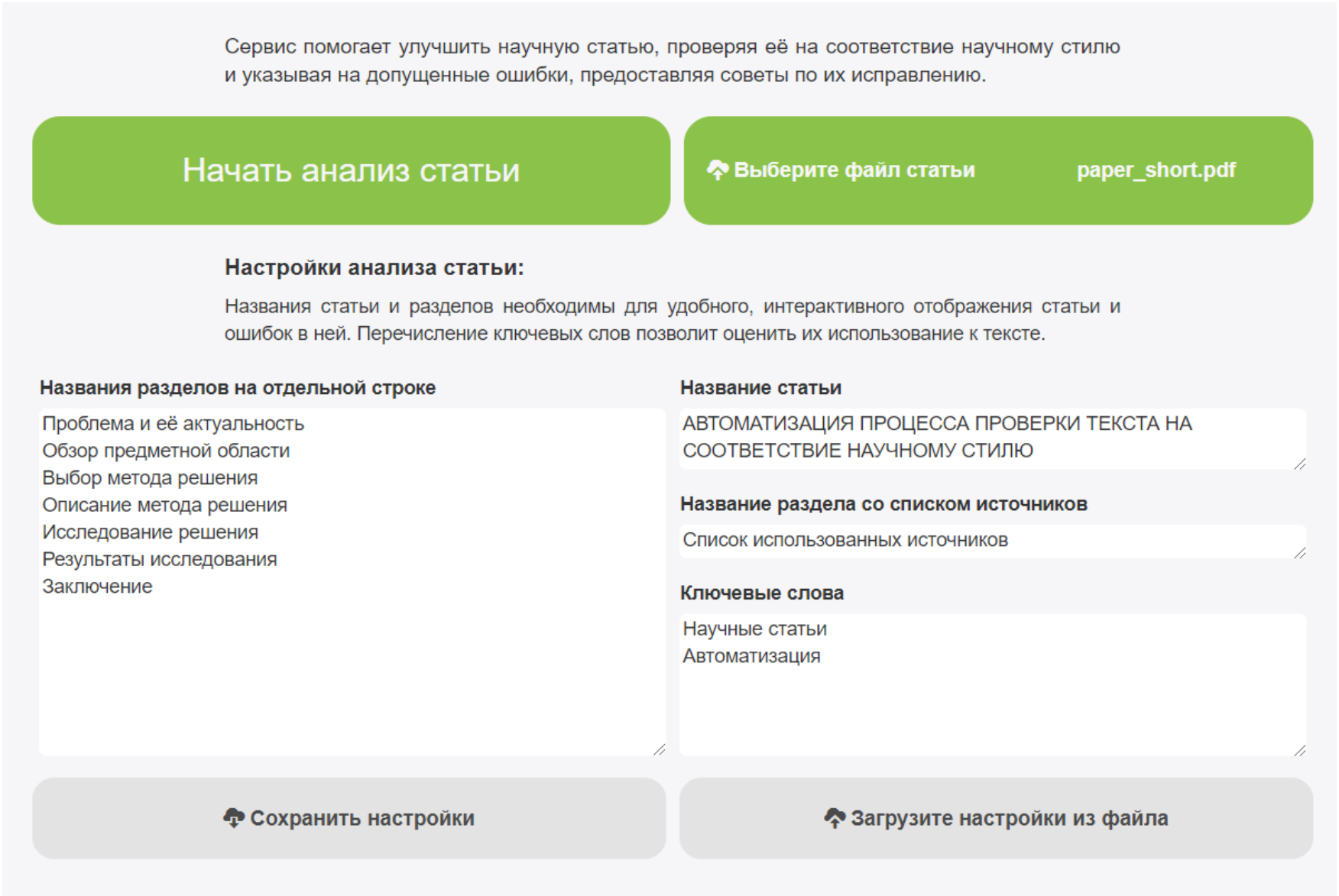
Стилистические:

- Использование личных местоимений;
- Использование обобщений;
- Необъективная оценка;
- Использование усилителей;
- Использование риторических вопросов.

Структурные:

- Отсутствие ссылки на указанный источник;
- Использование устаревшего источника;
- Отсутствие ссылки на рисунок;
- Отсутствие ссылки на таблицу;
- Наличие коротких разделов – разделов, состоящих менее чем из трёх предложений.
- Использование указанных ключевых слов в тексте.

Дополнительный слайд №6. Экран настройки анализа



Дополнительный слайд №7. Экран результата анализа статьи

Оценка стиля статьи:

39 из 100

Критерии:

Уровень водности

Процентное соотношение стоп-слов и общего количества слов в тексте

Значение: 23,820

Требования: Значение критерия должно находиться в интервале [14, 20]

Совет: Постарайтесь снизить количество используемых стоп-слов. Часто употребляемые стоп-слова в статье:
в: 47 раз
на: 23 раз
и: 22 раз
с: 15 раз
для: 14 раз
этот: 13 раз
он: 10 раз
который: 8 раз
к: 8 раз

Это тестовое предложение я добавил специально , оно содержит ошибки , которые точно должны быть выделены .

АВТОМАТИЗАЦИЯ ПРОЦЕССА ПРОВЕРКИ ТЕКСТА НА СООТВЕТСТВИЕ НАУЧНОМУ СТИЛЮ

Проблема и её актуальность

Соответствие статьи научному стилю является одним из основных критериев принятия статьи к публикации . В текущем виде , процесс проверки представляет собой отправку статьи на обзор третьим лицам , ожидание ответа , исправление недочетов и отправка на повторную проверку – это очень долго . В связи с этим , автоматизация данного процесса является актуальной задачей , позволяющей значительно ускорить процесс выявления ошибок для исправления , и в следствие этого ускорить сам процесс публикации статьи . В соответствие с этим возникает задача исследования возможности автоматизации процесса проверки научных статей на соответствие научному стилю . Также возникает необходимость предложить решение , позволяющее проверить научную статью по нескольким критериям , основываясь на проведенном исследовании .

Обзор предметной области

Научный стиль - наиболее строгий стиль речи , используемый для написания научных статей . Характеризуется использованием научной терминологии , исключая жаргонизмы . Научный стиль не допускает личного изложения [1] . Проверяя текст на соответствие научному стилю есть смысл реализовать и базовую проверку на качество текста . К такого рода анализу можно отнести SEO-анализ . SEO (search engine optimization) анализ [2-3] популярен и актуален в связи с необходимостью продвижения своих ресурсов , товаров и услуг в интернете . Основные термины SEO-анализа : Тошнота – это показатель повторений в текстовом документе ключевых слов и фраз . Синонимом тошноты является термин плотность [3] ; Стоп-слова – это слова в тексте , которые не несут смысловой нагрузки . Иначе их называют также шумовые слова [3] ; Вода - процентное соотношение стоп-слов и общего количества слов в тексте [3] . Уровень " воды " в тексте , его " тошнотность " и подсчет других числовых показателей , очевидно , можно автоматизировать . Но также важными показателями научной статьи являются её экспертность и полезность . На данный

Использование личного местоимения

Найдено ошибок: 11

Использование личных местоимений запрещено. Проверьте, можно ли удалить это местоимение без потери смысла.

Нет ссылки на источник

Найдено ошибок: 1

Необходимо хотя бы раз сослаться на каждый из перечисленных источников.

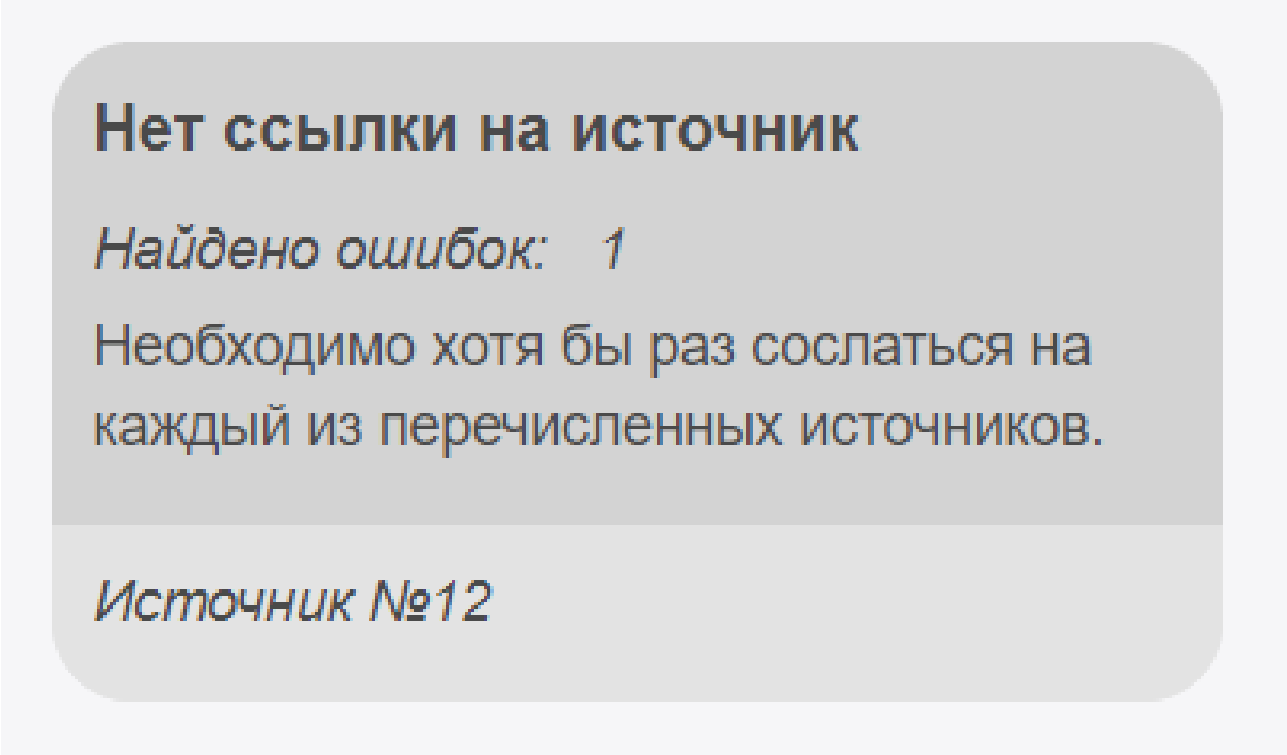
Короткий раздел

Найдено ошибок: 1

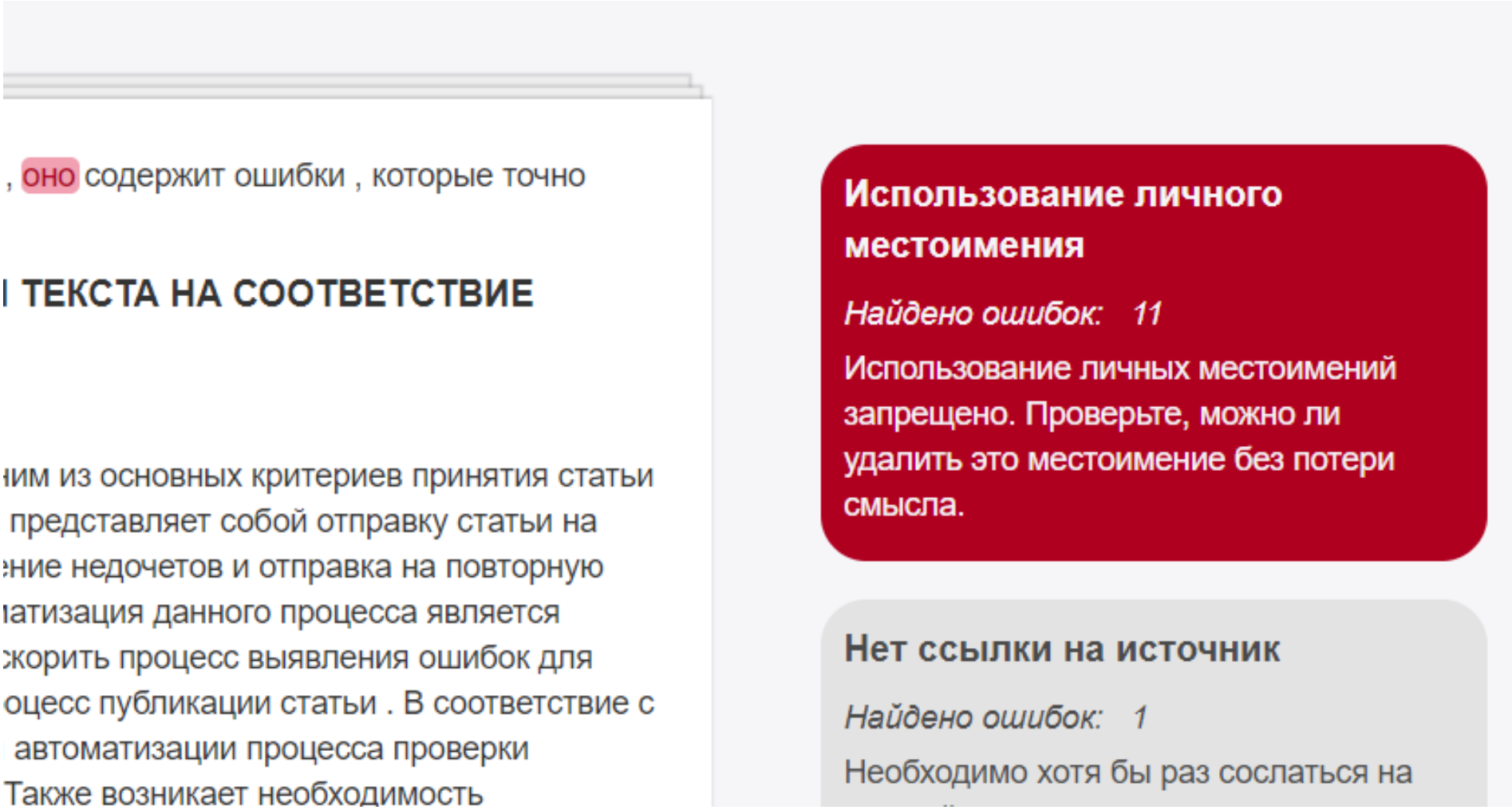
В разделе меньше трёх предложений. Постарайтесь расширить раздел, либо уберите его.

Дополнительный слайд №8. Пример отображения критерия проверки

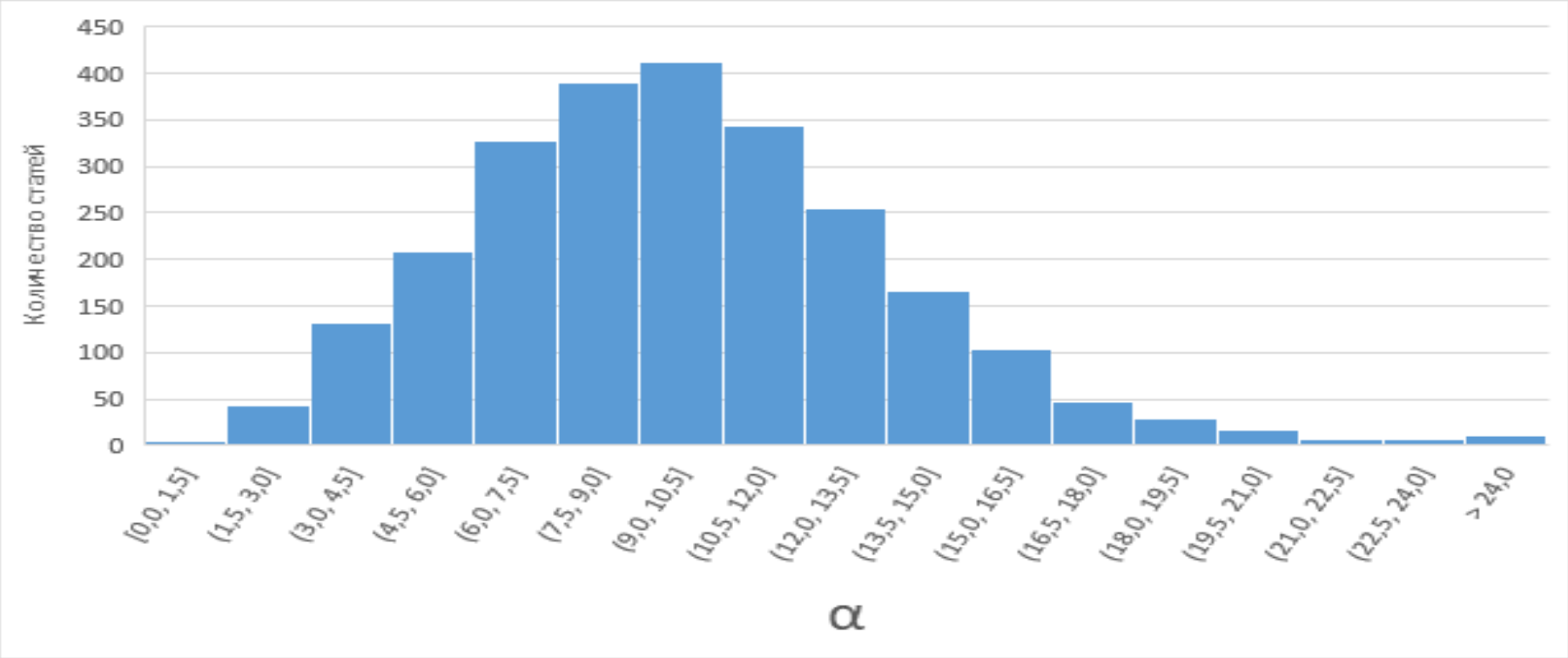
Дополнительный слайд №9. Пример отображения ошибки



Дополнительный слайд №10. Пример отображения выделения типа ошибки по слову

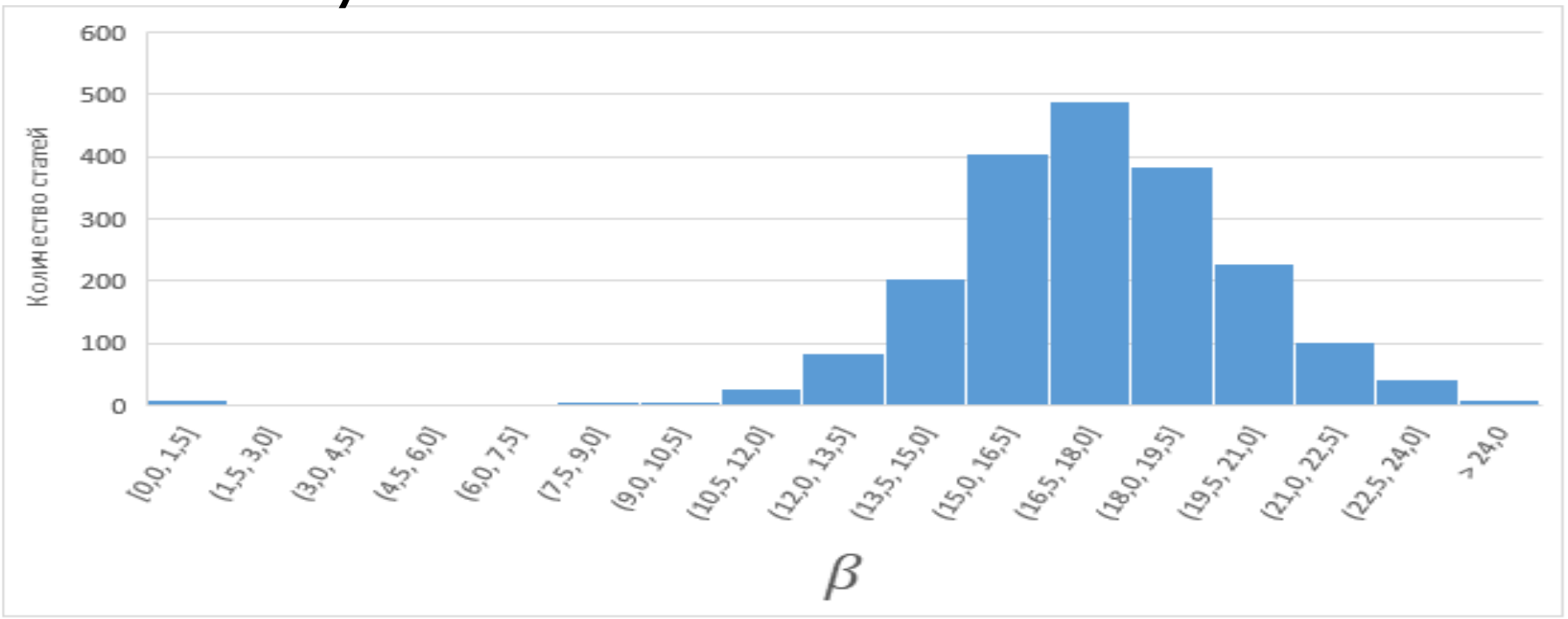


Дополнительный слайд №11. Полученные значения α по выборке



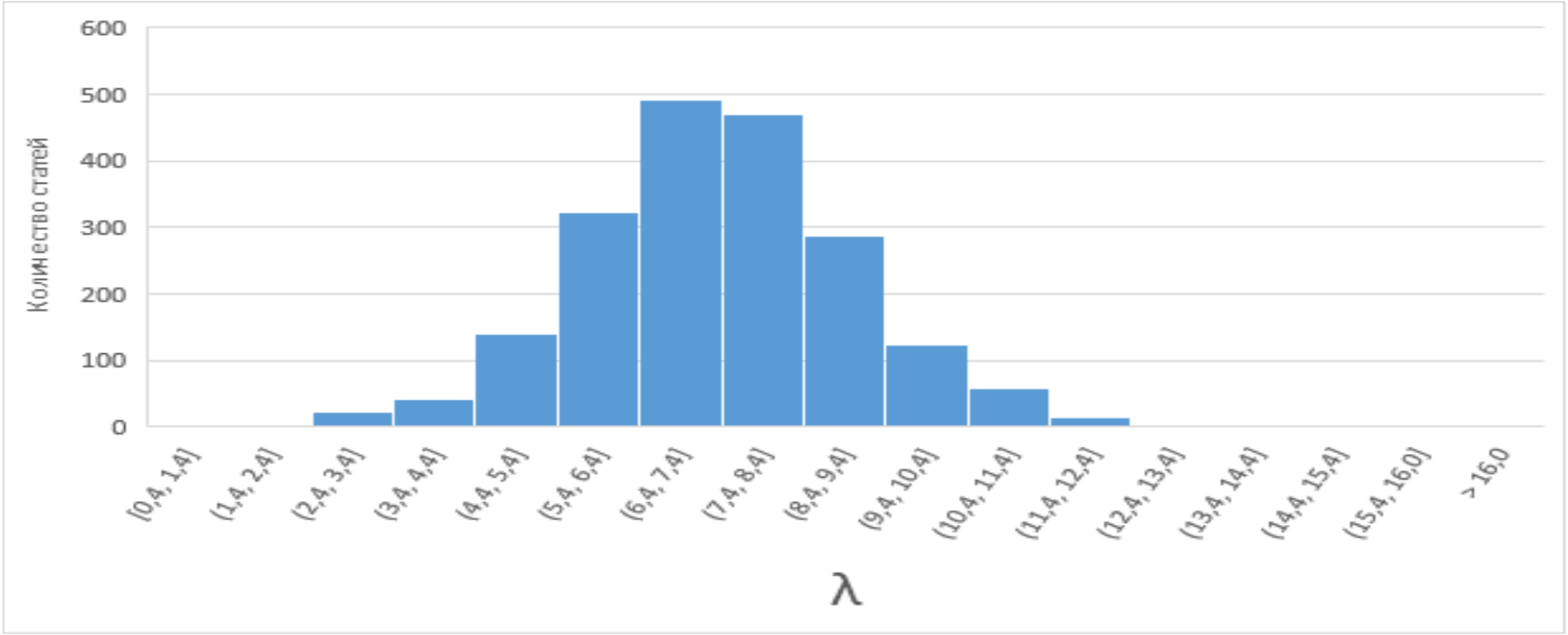
Выборка	Мат. ожидание	Дисперсия
α	9.822	3.902

Дополнительный слайд №12. Полученные значения β по выборке



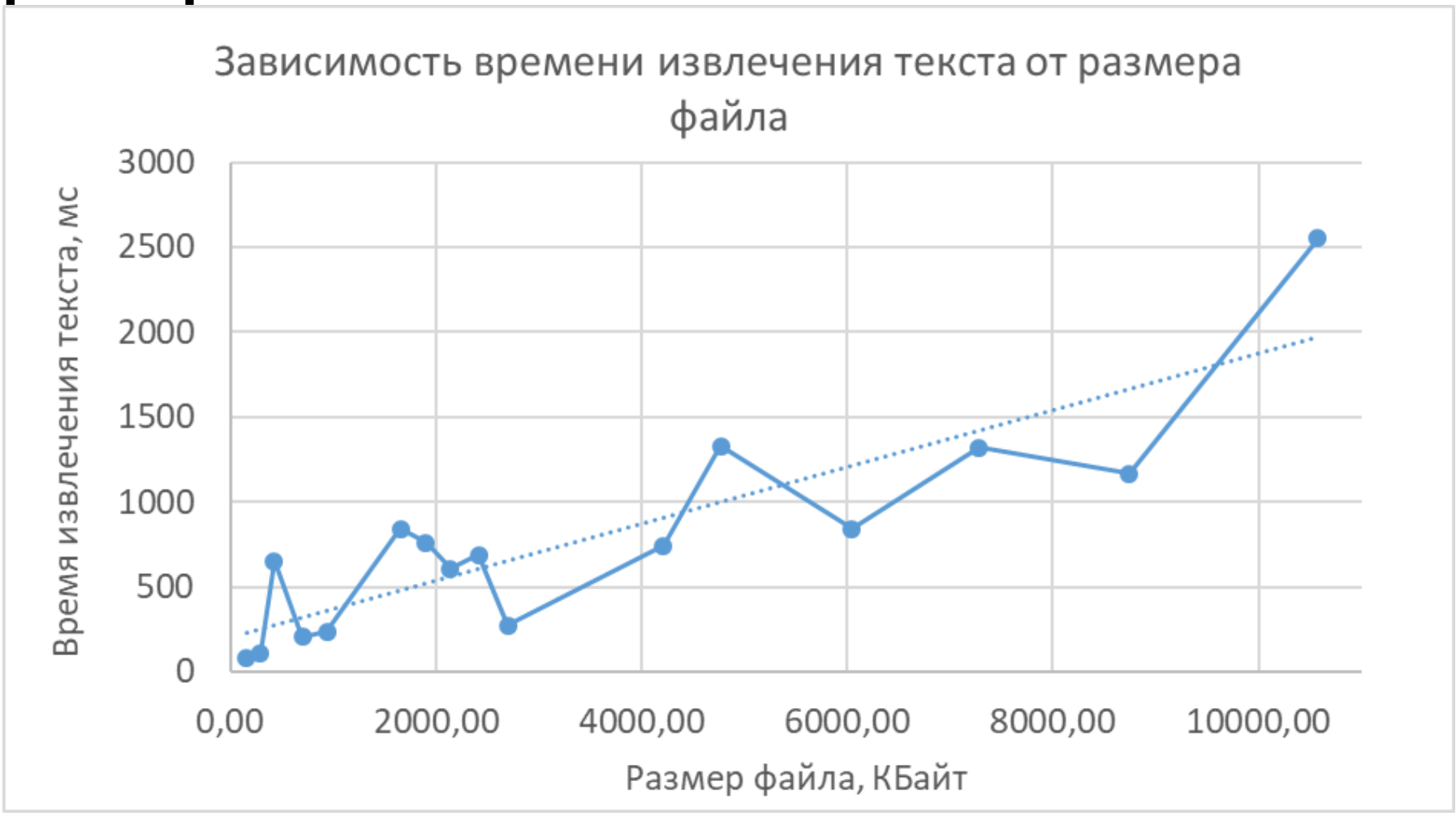
Выборка	Мат. ожидание	Дисперсия
β	17.145	3.082

Дополнительный слайд №13. Полученные значения λ по выборке



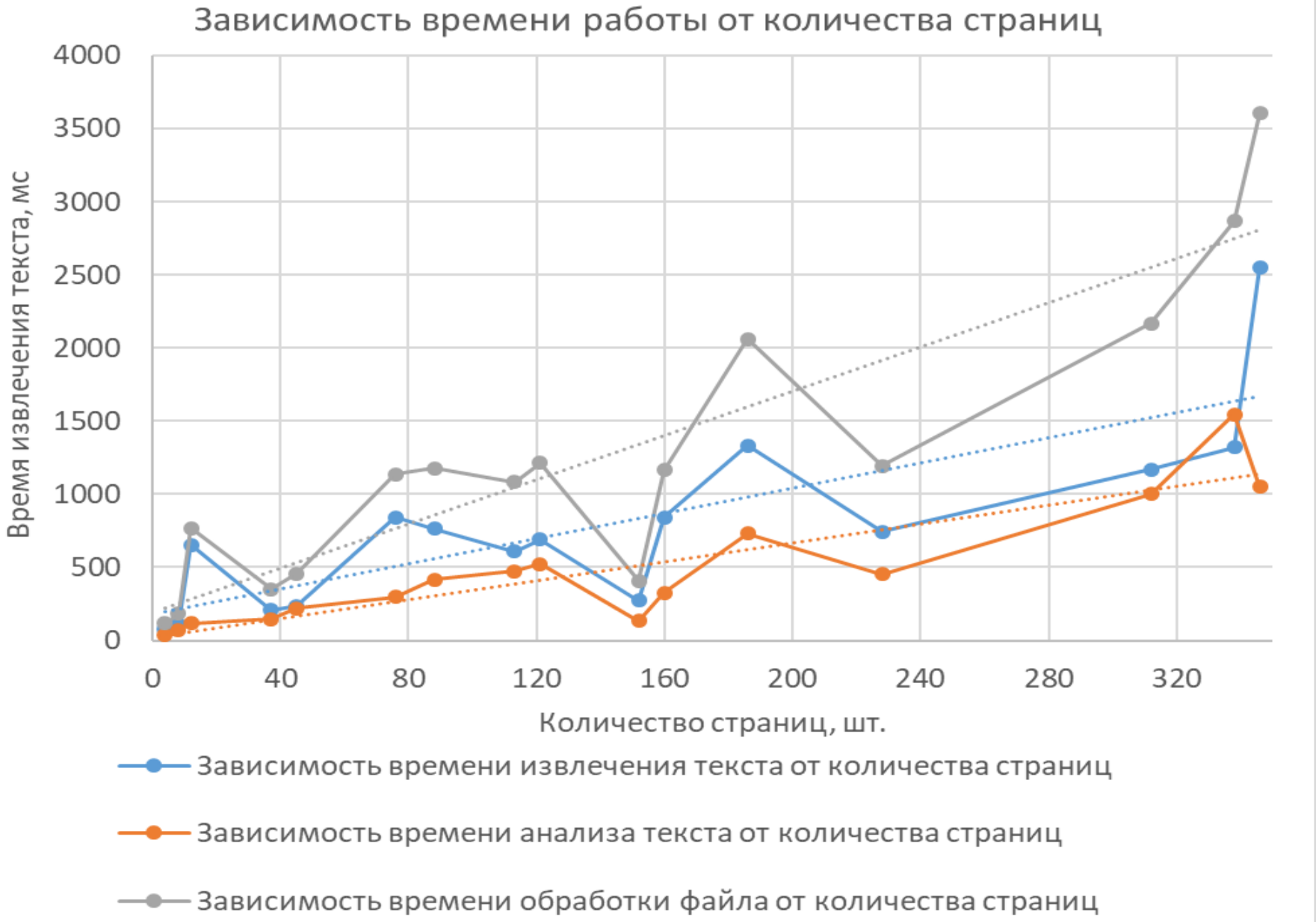
Выборка	Мат. ожидание	Дисперсия
λ	7.396	2.069

Дополнительный слайд №14. Зависимость времени извлечения текста из pdf файла от его размера



Дополнительный слайд №15. Зависимость времени анализа текста от количества символов

Дополнительный слайд №16



Дополнительный слайд №17. Оценка времени анализа статьи

$$T_E(x) = 0.1673x + 201.65$$

Где T_E – время извлечения текста в миллисекундах, x – размер файла в килобайтах.

$$T_A(y) = 0.0021y + 33.782$$

Где T_A – время анализа текста в миллисекундах, y – количество символов.

$$T(z) = 7.5639z + 194.49$$

Где T – время обработки файла в миллисекундах, z – количество страниц.

Дополнительный слайд №16. Потребление оперативной памяти приложением