

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 возможности автоматизации проверки статей на соответствие научному стилю

В результате обзора были выделены морфологические особенности научного стиля:

- **Запрет использования личных местоимений.** Личные и притяжательные местоимения (я, ты, мною, вы, наш) имеют отвлеченно-обобщенный характер и их употребление необходимо избегать;
- **Запрет использования неопределенных местоимений** (кое-что, что-нибудь). Эти местоимения, в силу их неопределенности, не употребляются.

Исследование возможности автоматизации проверки статей на соответствие научному стилю

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

Вводимые термины:

- Тошнота – это показатель повторений в текстовом документе ключевых слов и фраз. Синонимом тошноты является термин плотность.
- Стоп-слова – это слова в тексте, которые не несут смысловой нагрузки
- 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

Информационный стиль и SEO-анализ вводят морфологические ограничения:

- Использование слов усилителей (безусловно, очень, абсолютно и др.);
- Использование обобщений (со всего мира, весь, в общем);
- Необъективная оценка (уникальный, новейший);
- Использование риторических вопросов.

Исследование возможности автоматизации проверки статей на соответствие научному стилю

Автоматизируемые правила проверки научных статей в существующем курсе:

- Каждое ключевое слово упоминается в основном тексте хотя бы один раз;
- Более половины элементов списка литературы - актуальные и значимые научные работы;
- Все элементы списка литературы имеют минимум одно упоминание в тексте;
- Все рисунки и таблицы имеют подрисуночные подписи и ссылки в тексте.

Исследование возможности автоматизации проверки статей на соответствие научному стилю

Обзор аналогов

Аналог	Многокритерияльная 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 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:

$$K = B - \Phi$$

Где K – оценка статьи, B – базовое значение K ,
 Φ – штраф.

Математическая модель проверки статьи

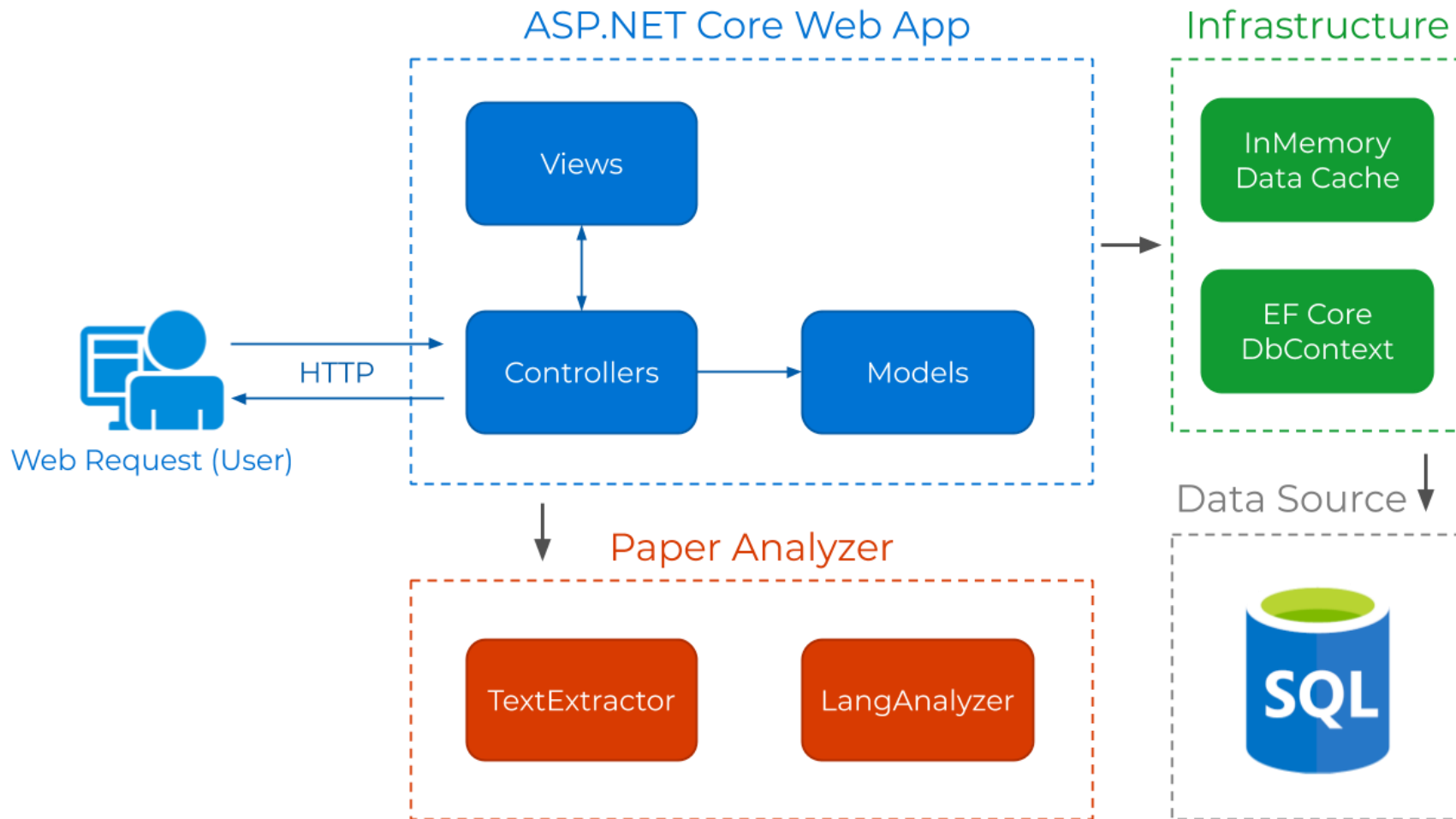
$$E(\alpha) = \begin{cases} 1, \alpha \in [6;14] \\ 0, \alpha \notin [6;14] \end{cases}$$

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

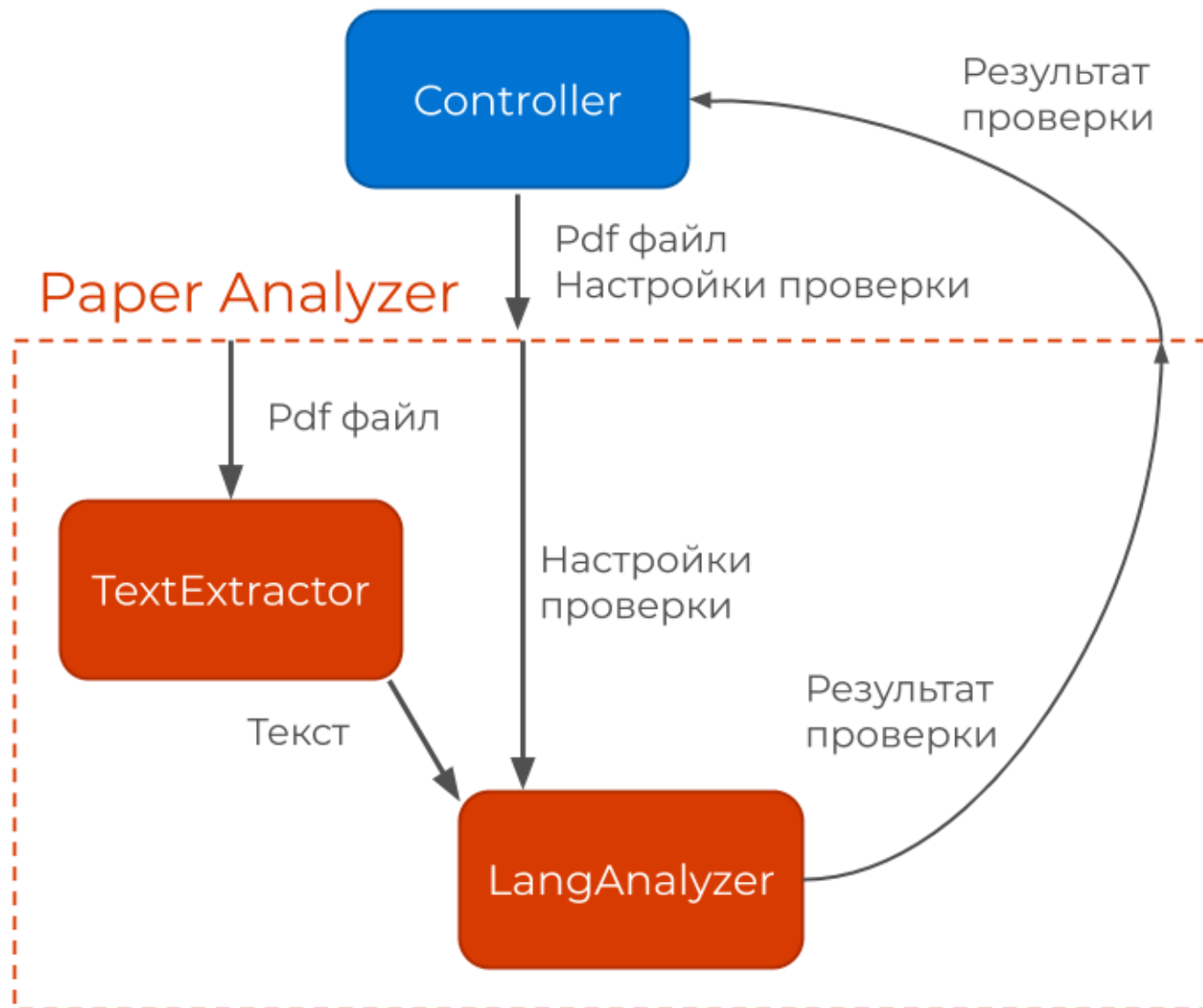
$$\Phi = C_4 \times N_1 + C_5 \times N_2$$

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

Разработанное решение. Архитектура



Разработанное решение. Алгоритм обработки



Заключение

- Было проведено исследование возможности автоматизации проверки научных статей на соответствие научному стилю, по результатам которого были выделены критерии проверки статей;
- Была построена математическая модель проверки статьи;
- Было проведено экспериментальное исследование на научных статьях, по результатам которого была настроена и формализована модель;
- Было проведено экспериментальное исследование на статьях и произведениях других жанров, показавшее корректность разработанной модели проверки;
- Было разработано решение в виде веб-сервиса.

Апробация работы

- Блеес Э.И., Заславский М.М., Андросов В.Ю. 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. Собственно-научный подстиль

В рамках данной работы была реализована проверка статей на соответствие собственно-научному подстилю. Собственно-научный подстиль — академическое изложение, адресованное специалистам.

Характеристики:

- Точность передаваемой информации;
- Убедительность аргументации;
- Логическая последовательность изложения;
- Лаконичность.

Цель подстиля — выявление и описание новых фактов, закономерностей, открытий.

Дополнительный слайд №2. 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 №3. 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 №4. Экран настройки анализа

Сервис помогает улучшить научную статью, проверяя её на соответствие научному стилю и указывая на допущенные ошибки, предоставляя советы по их исправлению.

Начать анализ статьи

Выберите файл статьи

paper_short.pdf

Настройки анализа статьи:

Названия статьи и разделов необходимы для удобного, интерактивного отображения статьи и ошибок в ней. Перечисление ключевых слов позволит оценить их использование к тексту.

Названия разделов на отдельной строке

Проблема и её актуальность
Обзор предметной области
Выбор метода решения
Описание метода решения
Исследование решения
Результаты исследования
Заключение

Название статьи

АВТОМАТИЗАЦИЯ ПРОЦЕССА ПРОВЕРКИ ТЕКСТА НА
СООТВЕТСТВИЕ НАУЧНОМУ СТИЛЮ

Название раздела со списком источников

Список использованных источников

Ключевые слова

Научные статьи
Автоматизация

Сохранить настройки

Загрузите настройки из файла

Дополнительный слайд №5. 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
по: 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 соответствии 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]. Проверка текста 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 ее 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 №6. Пример отображения критерия проверки

Тошнота



Показатель повторений в текстовом документе ключевых слов и фраз

Значение: 6,037

Требования: Значение критерия должно находиться в интервале [6, 14]

Дополнительный слайд №7. Пример отображения ошибки

Нет ссылки на источник

Найдено ошибок: 1

Необходимо хотя бы раз сослаться на каждый из перечисленных источников.

Источник №12

Дополнительный слайд №8. Пример отображения выделения типа ошибки по слову

, оно содержит ошибки , которые точно

ТЕКСТА НА СООТВЕТСТВИЕ

им из основных критериев принятия статьи
представляет собой отправку статьи на
ние недочетов и отправка на повторную
атизация данного процесса является
жорить процесс выявления ошибок для
оцесс публикации статьи . В соответствии с
автоматизации процесса проверки
Также возникает необходимость

Использование личного местоимения

Найдено ошибок: 11

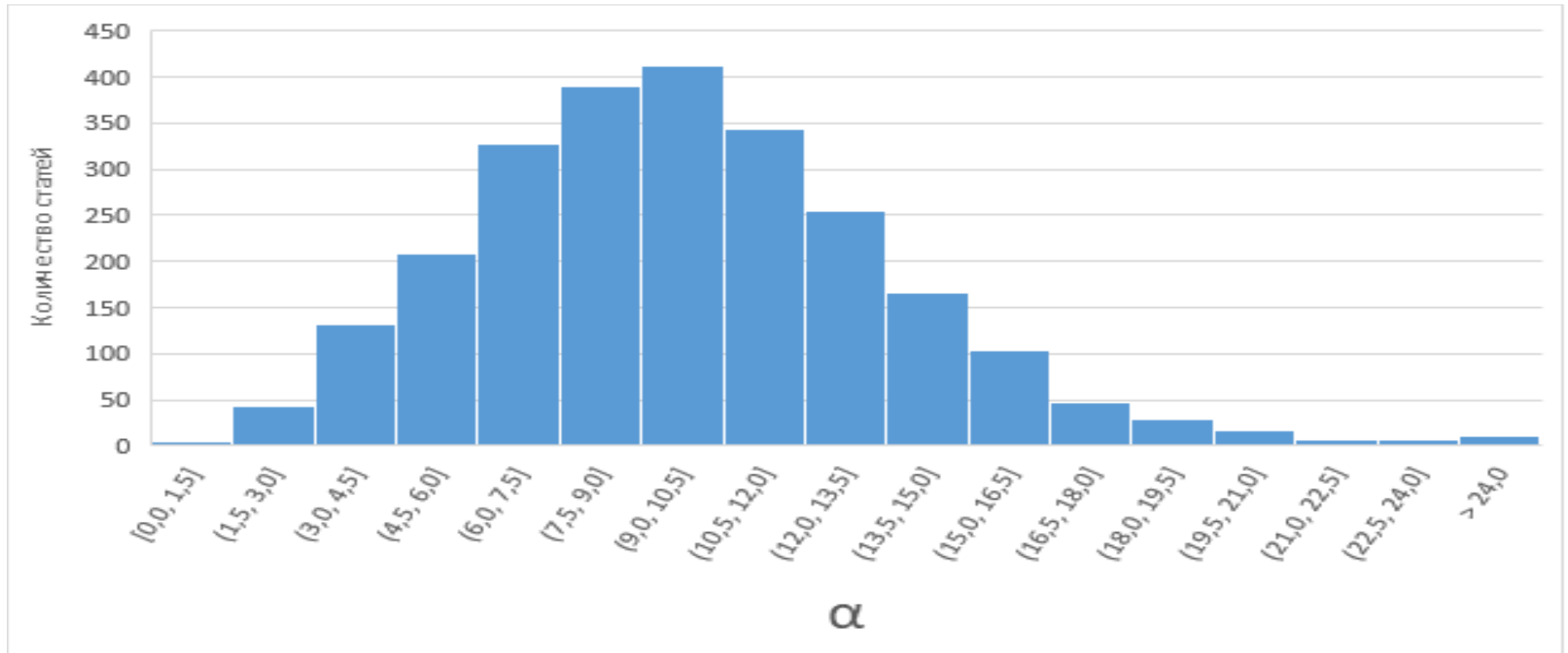
Использование личных местоимений запрещено. Проверьте, можно ли удалить это местоимение без потери смысла.

Нет ссылки на источник

Найдено ошибок: 1

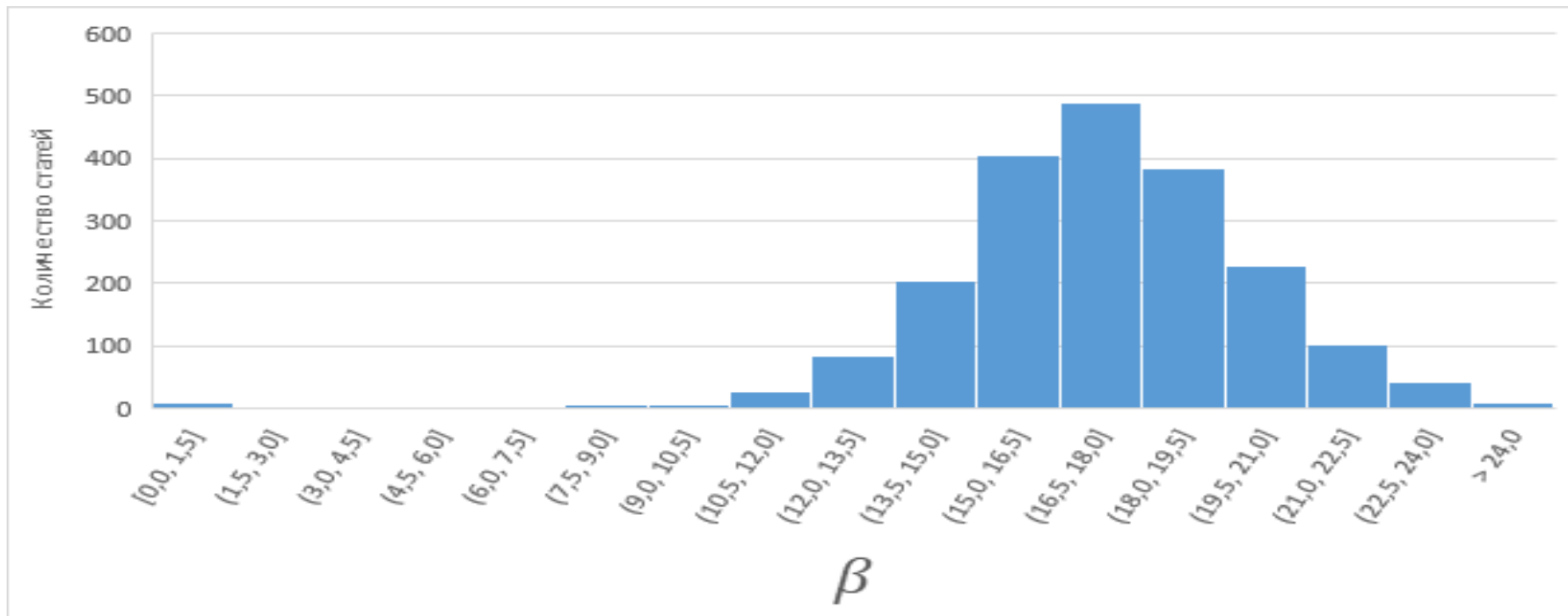
Необходимо хотя бы раз сослаться на

Дополнительный слайд №9. Полученные значения альфа по выборке



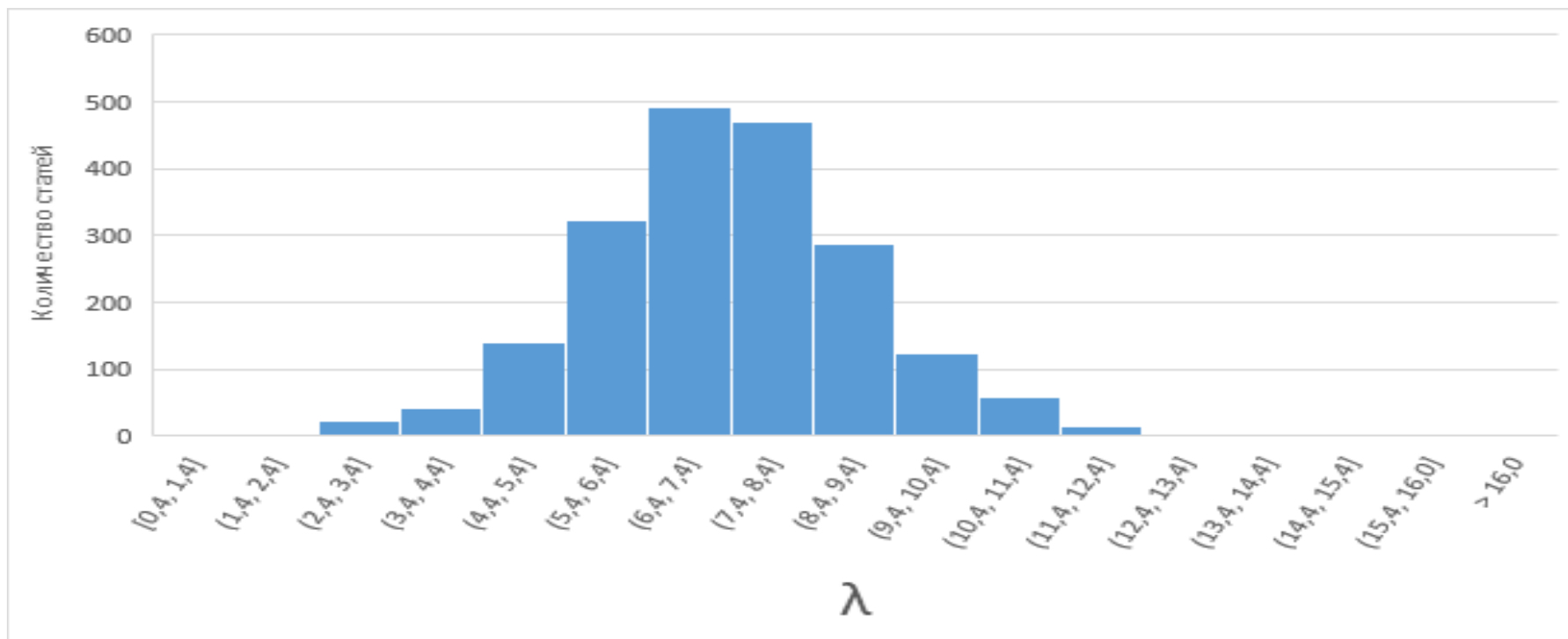
Выборка	Мат. ожидание	Дисперсия
α	9.822	3.902

Дополнительный слайд №9. Полученные значения бета по выборке



Выборка	Мат. ожидание	Дисперсия
β	17.145	3.082

Дополнительный слайд №9. Полученные значения лямбда по выборке



Выборка	Мат. ожидание	Дисперсия
λ	7.396	2.069