Экспериментальное исследование критериев соответствия текста научному стилю

Блеес Эдуард, студент 2 курса магистратуры, edw252@gmail.com
Марк Заславский, ассистент каф. МОЭВМ, mark.zaslavskiy@gmail.com

Проблема

Результатом предыдущей работы стало определение основных числовых критериев проверки статьи на соответствие научному стилю:

- Уровень ключевых слов в тексте – α,
- Процентное соотношение стоп-слов и общего количества слов в тексте – β,
- Значение отклонения текста статьи от идеальной кривой по Ципфу – λ.

Однако, для использования числовых критериев для оценки качества статьи, необходимо установить, как качество статьи связано со значениями этих числовых критериев.

Актуальность проблемы

- Статьи должны соответствовать научному стилю
- Статьи требуются бакалаврам, магистрам аспирантам, число которых растет

Цель и задачи работы

- Цель работы – необходимо исследовать взаимосвязь между качеством научного текста и значениями критериев α, β и λ
- Задачи – рассмотреть статистические свойства распределений значений критериев α, β и λ для научных статей и сформулировать необходимое, но не достаточное условие соответствия статьи научному стилю, проверить полученные условия на тестовой выборке.

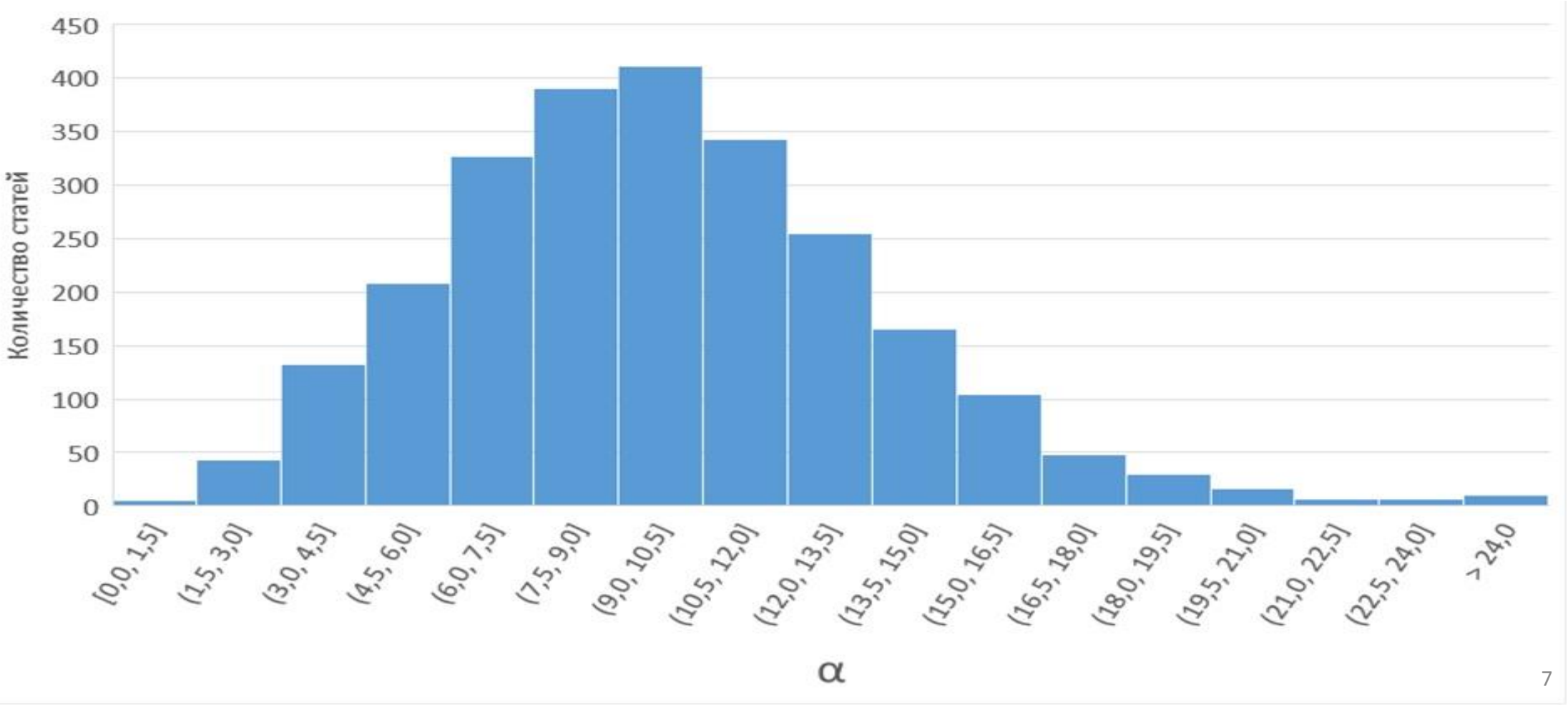
Область применения проекта

Сокращение времени публикации статей за счет автоматизации части процесса проверки статьи

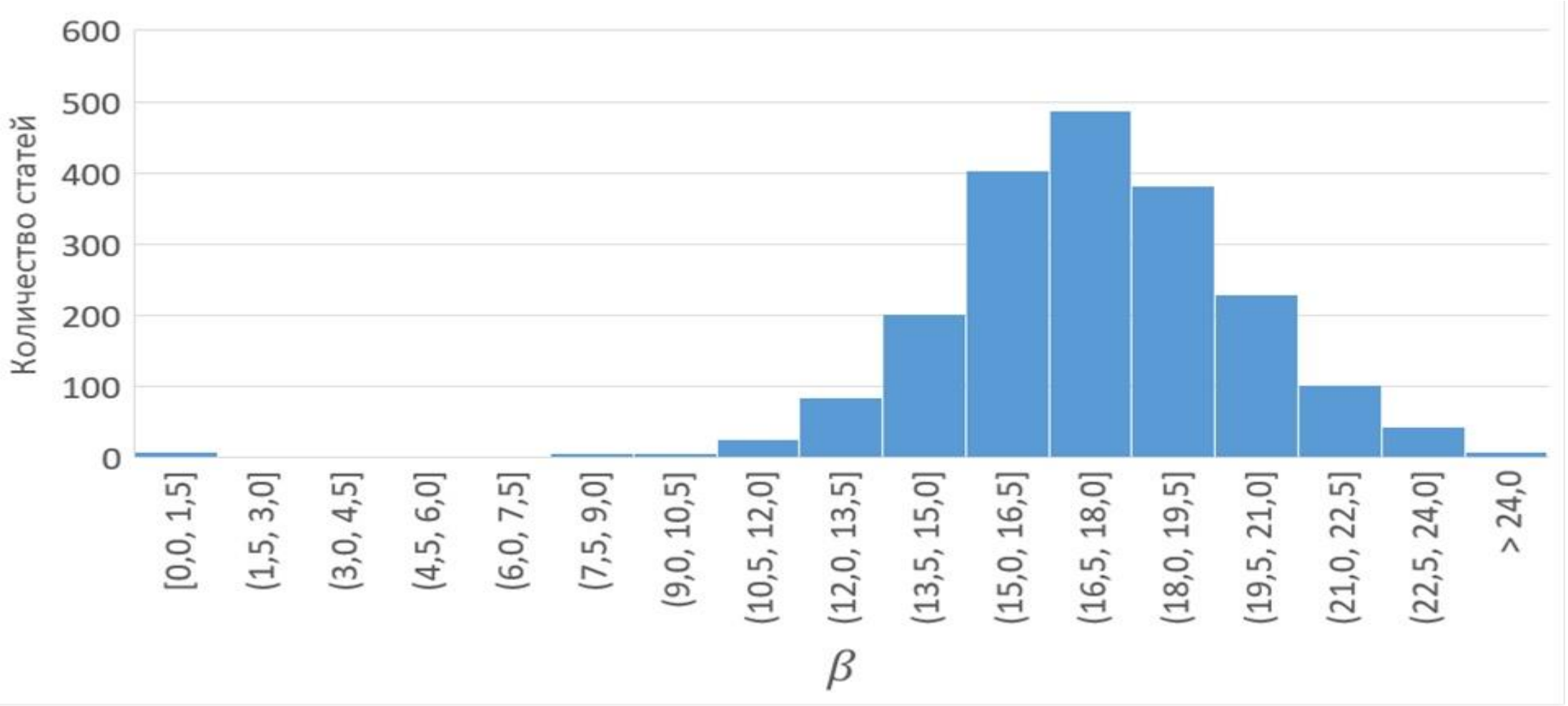
Решение

Было проведено исследование на выборке из 2500 статей опубликованных в ВАК и/или РИНЦ. В результате работы исполняемого сценария были получены значения числовых критериев по каждой из статей. После анализа результатов исполняемый сценарий был запущен на тестовой выборке, состоящей из бакалаврских работ студентов СПбГЭТУ "ЛЭТИ" 2016 и 2017 годов выпуска.

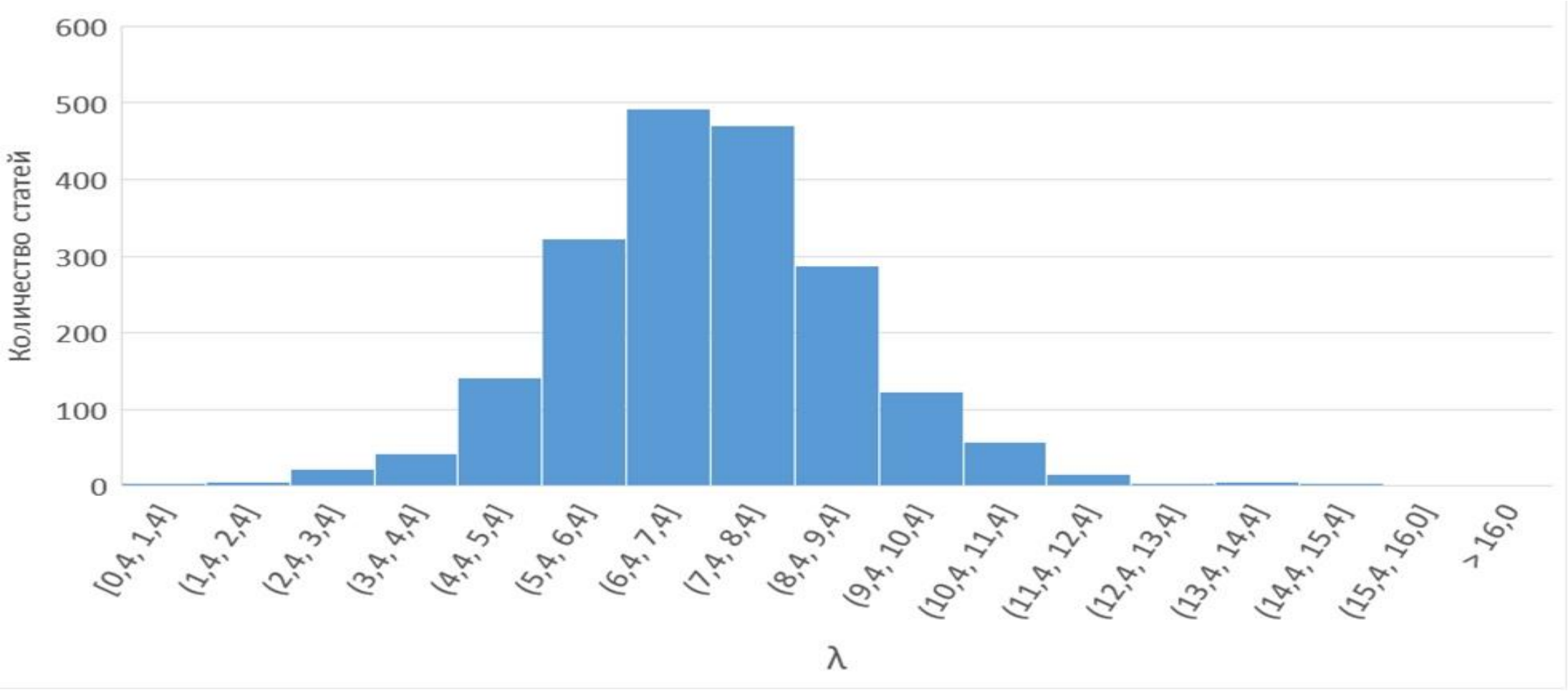
Исследование. Подчинение числовых критериев нормальному распределению. Критерий α



Исследование. Подчинение числовых критериев нормальному распределению. Критерий β



Исследование. Подчинение числовых критериев нормальному распределению. Критерий λ



Исследование. Анализ результатов

Нормальность распределений была показана с помощью трех тестов нормальности: критерий Шапиро-Уилка, критерий Колмогорова-Смирнов, критерий Андерсона-Дарлинга

Характеристики выборок

Выборка	Мат. ожидание	Дисперсия
α	9.822	3.902
β	17.145	3.082
λ	7.396	2.069

Установленные интервалы

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

Исследование независимости числовых критериев

Коэффициент корреляции Пирсона:

$$r_{XY} = \frac{\operatorname{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 (Y - \bar{Y})^2}}$$

Полученная матрица корреляций:

$$\begin{pmatrix} 1 & -0.07 & 0.22 \\ -0.07 & 1 & 0.01 \\ 0.22 & 0.01 & 1 \end{pmatrix}$$

Запуски на тестовой выборке и других текстах

Перед сравнением примем следующие условия оценки работ с помощью анализа критериев:

Оценка	Количество критериев, попадающих в интервал
5	$N \in [2, 3]$
4	$N \in [1, 2]$
3	$N \in [0, 1]$

Приняли допущение, что оценка за дипломную работу отражает его качество, несмотря на то что на самом деле, на оценку влияет множество других параметров.

В ходе проверки статей было выявлено 28 ошибок 1 или 2 рода, то есть в 65% случаев оценка по анализу критериев совпала с оценкой, поставленной аттестационной комиссией.

Запуски на тестовой выборке и других текстах

Текст	α	α ∈ [6, 14]	β	β ∈ [14, 20]	λ	λ ∈ [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
Интернет-статья «Наши с вами персональные данные ничего не стоят»	10.56	Да	32.10	Нет	6.84	Да
Интернет-статья «Рассказ о том, как я ворую номера кредиток и пароли у посетителей ваших сайтов»	6.61	Да	36.46	Нет	6.82	Да
Интернет-статья «Трехмерный движок на формулах Excel для чайников»	11.61	Да	27.91	Нет	9.27	Да
«Капитал» Карла Маркса	5.84	Нет	28.94	Нет	138.22	Нет
«Идиот» Фёдора Достоевского	6.65	Да	45.65	Нет	53.12	Нет
«Мёртвые души» Николая Гоголя	7.14	Да	40.81	Нет	35.58	Нет
«Путешествие к центру Земли» Жюля Верна	5.03	Нет	35.19	Нет	21.56	Нет

Сформулированное правило

$$\alpha \in [6.0;\ 14.0] \wedge \beta \in [14.0,\ 20.0] \wedge \lambda \in [5.5,\ 9.5]$$

Стиль статьи можно считать научным, если по всем трём показателям, значения статьи попадают в заданные интервалы.

Планы по развитию решения

Создание Web-приложения для удобной проверки статьи на соответствие научному стилю

Спасибо за внимание

Репозиторий: https://github.com/EduardBlees/Master-s-thesis

Предыдущая работа: 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. 24 Апреля 2018 Научно-исследовательский корпус СПбПУ; 2 учебный корпус, Политехническая ул., д.29. 2018. - С. 118-121;

Блеес Эдуард, студент 2 курса магистратуры, edw252@gmail.com

Марк Заславский, ассистент каф. МОЭВМ, mark.zaslavskiy@gmail.com