Название статьи

Введение

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 – данные этапы могут занимать достаточно много времени. 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, разумно 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 сети Интернет. SEO анализ текста дает возможность понять, не переспамлен ли текст, насколько велика его тошнота, или не преобладает ли в нем вода, какие слова являются подавляющими и т.д.
Более подробный обзор предметной области приведен в предыдущей статье [ССЫЛКА].

Проблема

Результатом предыдущей работы стало определение основных числовых критериев проверки статьи на соответствие научному стилю. Однако, исследование критериев было недостаточным, в связи с чем возникла необходимость более подробного исследования для четкой формулировки числовых критериев проверки.

Решение

Принято решение запустить исполняемый сценарий на более крупной выборке научных статей для дальнейшего анализа полученных значений числовых критериев с целью формулирования оправданных критериев оценки научных работ по этим критериям. Было проведено исследование на выборке из 2500 статей опубликованных в ВАК и/или RSCI. 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.

Получение выборки статей

Выборка из 2500 статей была получена с помощью другого исполняемого сценария, который выполняет веб-скрэпинг научной интернет-библиотеки "Киберленика"

[ССЫЛКА]. Веб-скрэпинг - техника получения данных из человеко-читаемых данных, размещенных в веб-ресурсах. [ССЫЛКА]
Были загружены статьи, опубликованные в ВАК и/или RSCI, в разделах библиотеки "Информатика" и "Автоматика и вычислительная техника" для того, чтобы определить значения числовых критериев проверки научных статей технической направленности.

Исследование

Подчинение числовых критериев нормальному распределению

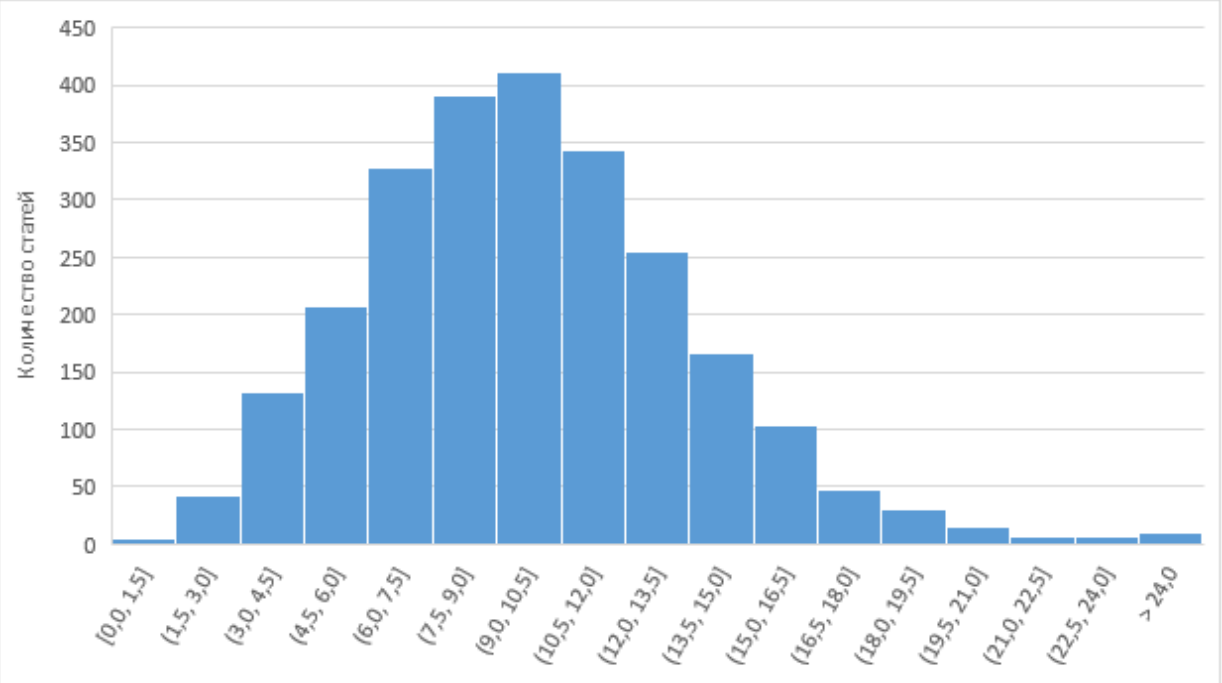

Рисунок 1 – Гистограмма распределения значений уровня ключевых слов в тексте статей из выборки

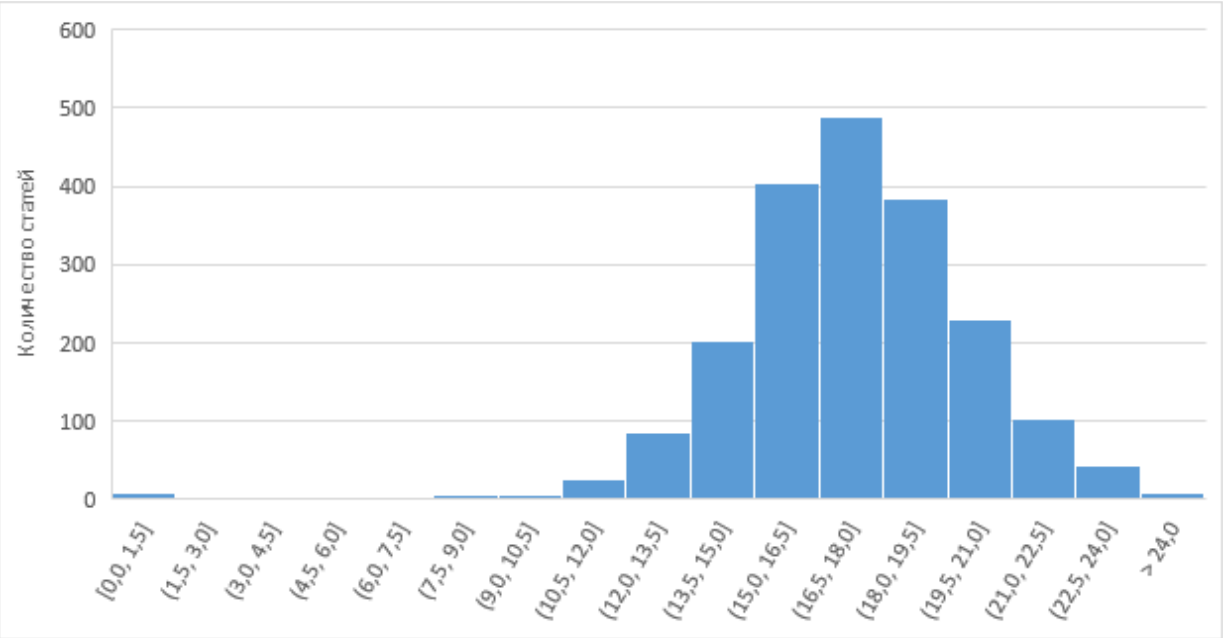

Рисунок 2 – Гистограмма распределения значений уровня водности текста статей из выборки

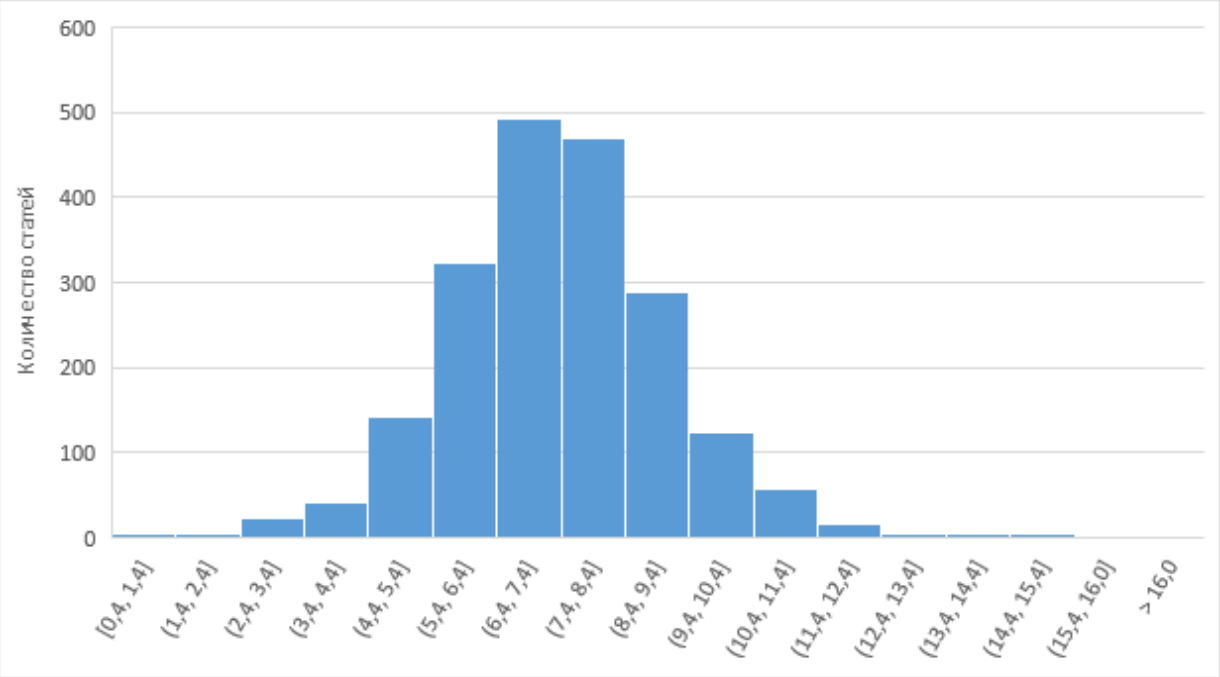

Рисунок 3 – Гистограмма распределения значений отклонения от идеальной кривой по Ципфу текста статей из выборки

Из рис. 1-3 видно, что распределения похожи на нормальные. Для доказательства воспользуемся тремя тестами нормальности: критерий Шапиро-Уилка [ССЫЛКА], критерий Колмогорова [ССЫЛКА], критерий Андерсона [ССЫЛКА]. В каждом из тестов проверяется нулевая гипотеза [ССЫЛКА], о том, что каждая выборка получена из нормального распределения. Так, нулевая гипотеза считается верной до того момента, пока нельзя доказать обратное. Статистическая значимость для тестов равна 0,05. P-значение — величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы.

Использовалась реализация тестов из статистической библиотеки SciPy [ССЫЛКА]. На выходе каждый тест выдает два значения – tst-statistics и P-value. В случае, если значение p-value близко к 0, или значительно меньше test-statistics – нулевая гипотеза не может быть отвергнута.

Результаты по каждому числовому критерию представлены в табл. 1-3:

Таблица 1 - результаты тестов для выборки значений уровня ключевых слов в тексте

Критерий	test-statistics	p-value
Шапиро	0.967	1.407e-23
Колмогоров	0.309	0.0
Андерсон	8.293	0.786

Таблица 2 - результаты тестов для выборки значений водности текста

Критерий	test-statistics	p-value
Шапиро	0.942	3.815e-30
Колмогоров	0.229	0.0
Андерсон	14.957	0.786

Таблица 3 - результаты тестов для выборки значений отклонения текста от идеальнйо кривой по Ципфу

Критерий	test-statistics	p-value
Шапиро	0.864	3.512e-42
Колмогоров	0.129	0.0
Андерсон	28.732	0.786

Как видно из результатов тестов – нет поводов отклонить нулевую гипотезу для каждой выборки, т.е. можно считать, что каждый числовой критерий подчиняется нормальному закону распределения.

В таблице 4 представлены математическое ожидание и дисперсия каждой из выборок:

Таблица 4 – Характеристики выборок

Выборка	Мат. ожидание	Дисперсия
Тошнота	9.822	3.902
Водность	17.145	3.082
Отклонение от идеальной кривой по Ципфу	7.396	2.069

На основе этих данных были установлены интервалы для каждого из числовых критериев:

Таблица 5 – Установленные интервалы

Критерий	Интервал
Тошнота	~ [6, 14]
Водность	~ [14, 20]
Отклонение от идеальной кривой по Ципфу	~ [5.5, 9.5]

Независимость числовых критериев

Независимость числовых критериев друг от друга показывает ценность каждого из них в отдельности – ни один из критериев не дублирует уже известную информацию [НЕ знаю как нормально сказать ИСПРАВИТЬ]. Для доказательства этого была построена матрица ковариации. Был использован линейный коэффициент корреляции (коэффициент корреляции Пирсона) для расчета корреляции числовых критериев на основе полученных выборок:

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

Рисунок 4 - Формула линейного коэффициента корреляции

Полученная матрица ковариации:

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

Коэффициенты корреляции близки к 0 или незначительны, следовательно, числовые критерии независимы.

Запуски на тестовой выборке и корчевателе

?

Выводы

* Была увеличена выборка, применены мат методы, получили "хорошие интервалы"

* на этом все с численными параметрами, в будущем планируется сделать веб сервис с проверками типовых ошибок и тд