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 рецензирование, 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.

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, следует в первую очередь 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 ресурсов, товаров и услуг в сети Интернет. SEO анализ текста дает возможность понять, насколько часто употребляются ключевые слова в тексте, как много в тексте слов, не имеющих смысловой нагрузки и т.д.
SEO-анализе вводит следующие термины для двух критериев, которые проверяются в данной работе:
Тошнота – это показатель повторений в текстовом документе ключевых слов и фраз. Синонимом тошноты является термин плотность [2]. Вода - процентное соотношение стоп-слов и общего количества слов в тексте [2]. Так как эти критерии вычисляемы, то можно автоматизировать их получение.
Так же существует эмпирическая закономерность распределения частоты слов естественного языка - Закон Ципфа: если все слова языка или достаточно длинного текста упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n [8-9]. Соответствие распределения слов в тексте закону Ципфа говорит об уровне его естественности. Рассчет этого критерия так же можно автоматизировать.
Также важными показателями научной статьи являются её экспертность и полезность. На данный момент это может проверить только специалист в данной области, но разработки подобных инструментов ведутся [ССЫЛКА].
Более подробный обзор предметной области и вычисления критериев приведен в предыдущей статье [ССЫЛКА].

Проблема

Результатом предыдущей работы стало определение основных числовых критериев проверки статьи на соответствие научному стилю. Для удобства обозначим данные критерии:

- Тошнота или уровень ключевых слов в тексте – α ,

- Уровень воды в тексте или процентное соотношение стоп-слов и общего количества слов в тексте – β,
- Значение отклонения текста статьи от идеальной кривой по Ципфу [ССЫЛКА на предыдущую работу] – λ.

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

Решение

Принято решение запустить исполняемый сценарий, разработанный ранее и описанных в предыдущей работе [ССЫЛКА на пред. работу] на более крупной выборке научных статей для дальнейшего анализа полученных значений числовых критериев с целью формулирования оправданных критериев оценки научных работ по этим показателям. Было проведено исследование на выборке из 2500 статей опубликованных в ВАК [ССЫЛКА] и/или РИНЦ [ССЫЛКА]. 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 [ССЫЛКА], который выполняет веб-скрэпинг [ССЫЛКА] научной интернет-библиотеки "Киберленика" [ССЫЛКА]. Были загружены статьи технической направленности, опубликованные в ВАК и/или РИНЦ.

Исследование

В рамках исследования проверялась гипотеза о том, что качество научной статьи влияет на значения ранее определенных числовых критериев, а также то, что полученная выборка значений будет соответствовать нормальному распределению.
Исследование на выборке из 2500 прошедших рецензирование и опубликованных статей позволит получить математические параметры распределений, что позволит установить пороговые значения числовых критериев для статей хорошего качества.

Подчинение числовых критериев нормальному распределению

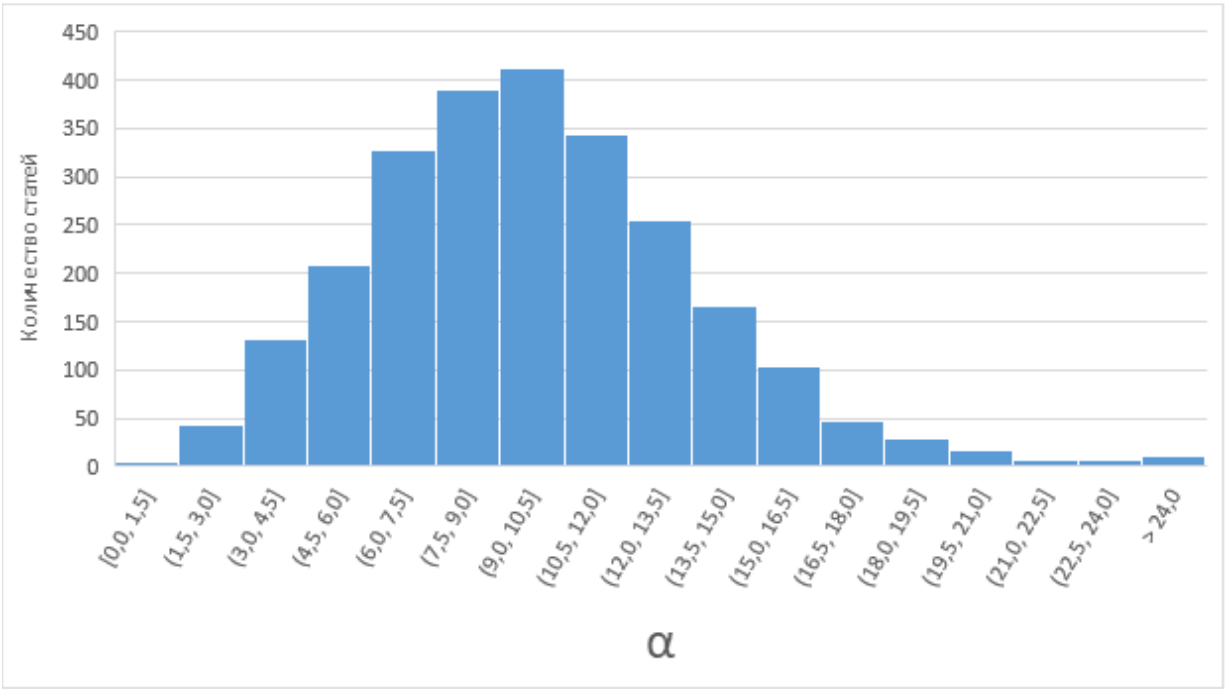

Рисунок 1 – Гистограмма распределения значений уровня ключевых слов в тексте статей из выборки

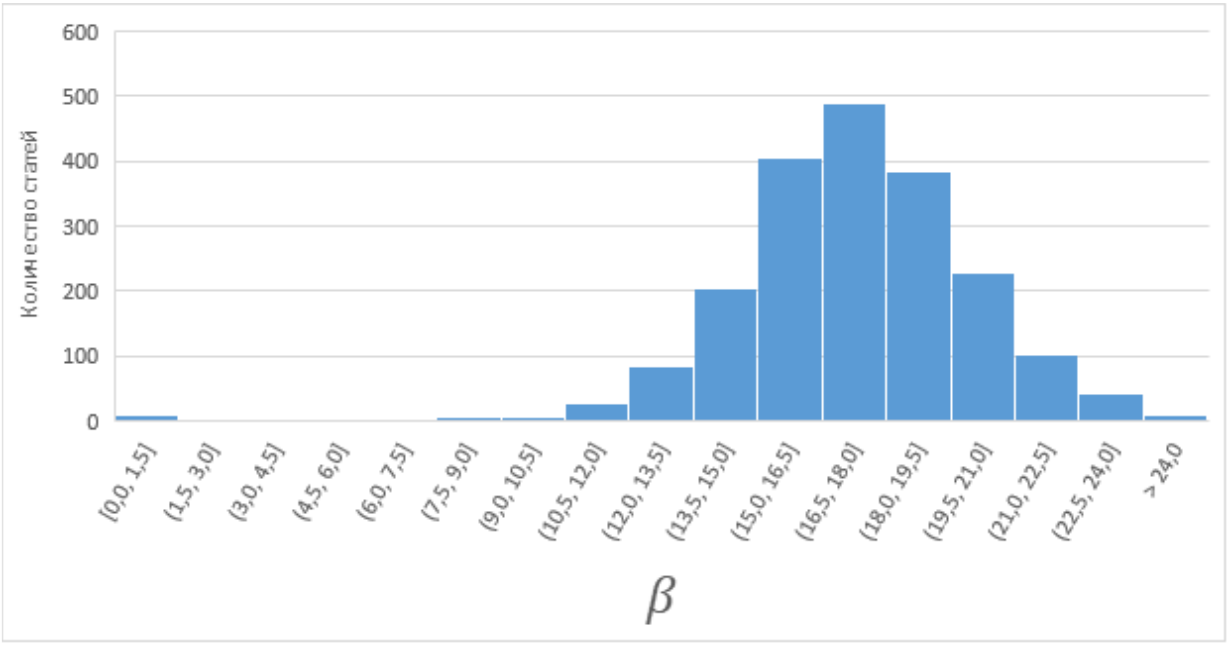

Рисунок 2 – Гистограмма распределения значений уровня водности текста статей из выборки

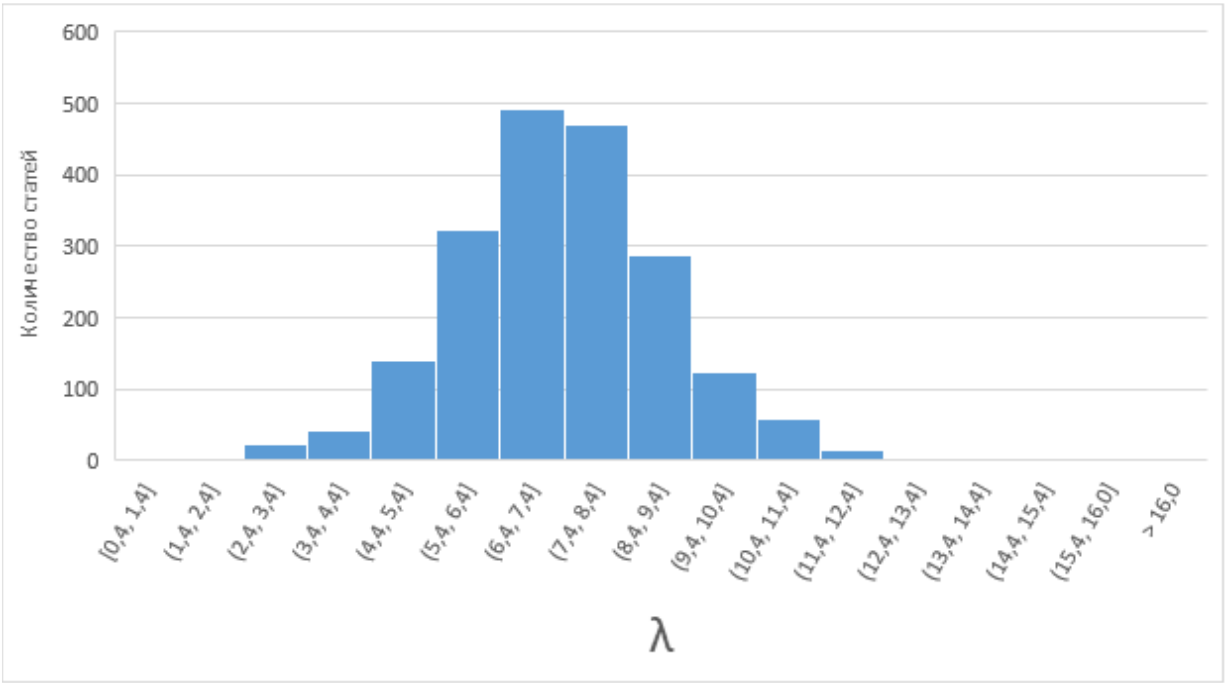

Рисунок 3 – Гистограмма распределения значений отклонения от идеальной кривой по Ципфу текста статей из выборки

Из рис. 1-3 видно, что у каждого из распределений наблюдается четкий пик и большинство значений сконцентрированы вокруг него симметрично, в связи с чем можно предположить, что распределения 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 [ССЫЛКА] для тестов равна 0,05. P-значение [ССЫЛКА] — величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы.

Использовалась [ССЫЛКА на скрипт] реализация тестов из статистической библиотеки SciPy [ССЫЛКА]. На выходе каждый тест выдает два значения – D (Статистика критерия для эмпирической функции распределения [ССЫЛКА]) и P-значение. В случае, если значение P-значение близко к 0, или значительно меньше D – 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	D	P-значение
Шапиро	9.67e-1	1.407e-23
Колмогоров	3.09e-1	0.0
Андерсон	8.293	7.866e-1

Таблица 2 - результаты тестов для выборки значений водности текста

Критерий	test-statistics	p-value
Шапиро	9.42e-1	3.815e-30
Колмогоров	2.29e-1	0.0

Андерсон	1.4957e+1	7.866e-1

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

Критерий	test-statistics	p-value
Шапиро	8.64e-1	3.512e-42
Колмогоров	1.29e-1	0.0
Андерсон	2.8732e+1	7.866e-1

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
α	9.822	3.902
β	17.145	3.082
λ	7.396	2.069

Так как распределения можно считать нормальными, то, согласно эмпирическому правилу [https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule – ПОМЕНЯТЬ НЕ НА ВИКИПЕДИЮ], более 2/3 распределения будет содержаться в следующем интервале $[\mu-\sigma, \mu+\sigma]$, где μ – среднее значение выборки, а σ – среднеквадратичное отклонение.

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

Таблица 5 – Установленные интервалы

Критерий	Интервал
α	~ [6, 14]
β	~ [14, 20]
λ	~ [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. 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, где [СДЕЛАТЬ]

Полученная матрица ковариации:

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

Коэффициент корреляции Пирсона может принимать значения от -1 до 1, где 0 означает полную независимость переменных друг от друга. Полученный коэффициент корреляции между α и β равен -0.07, а между β и λ равен 0.01, что позволяет утверждать о независимости данных критериев. Между критериями α и λ наблюдается небольшая зависимость.

Запуски на тестовой выборке и корчевателе

Для проверки адекватности полученных критериев, были использованы 80 дипломных бакалаврских работ студентов СПбГЭТУ «ЛЭТИ» кафедры МОЭВМ 2016 и 2017 годов. Кафедрой были предоставлены оценки данных работ, что позволит сравнить их с результатами анализа критериев, и подсчитать количество ошибок 1 и 2 рода [ССЫЛКА]. Перед сравнением примем следующие условия оценки работ с помощью анализа критериев:

Оценка	Количество критериев, попадающих в интервал
5	2-3 / 3
4	1-2 / 3
3	0-1 / 3

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. Данный эксперимент нельзя считать точным в связи с множеством критериев, которые влияют на оценку бакалаврской работы.
Так же была оценена работа «Корчеватель» [ССЫЛКА] – сгенерированная в научном стиле, не имеющая смысла статья, которая была принята для публикации в различные научные издания. По результатам проверки [ССЫЛКА на результаты в репозитории], значения всех трёх критериев попали в установленные интервалы, т.е. работу можно считать «отличной», так решили и рецензенты научных изданий. Но это снова подчеркнуло факт того, что данные критерии не проверяют содержание и полезность статьи.

Выводы

В результате работы были сформулированы три числовых критерия проверки статьи на соответствие научному стилю, были установлены пороговые значения данных критериев, позволяющие оценивать качество статей.

В дальнейшем планируется разработать веб-сервис, выполняющий проверку статей по полученным критериям, а так же анализирующий форматирование статей и семантику на соответствие научному стилю [ССЫЛКА на учебник по русскому о научном стиле, уже ссылался].