Экспериментальное исследование критериев соответствия текста научному стилю

Аннотация

В данной статье приведено экспериментальное исследование критериев соответствия текста научному стилю. Были исследованы 3 числовых критерия, полученных в предыдущей работе, в которой были рассмотрены возможности автоматизации процесса проверки научных статей на соответствие научному стилю, и в результате которой было показано, что часть критериев проверки может быть автоматизирована. Исследование проводилось с использованием исполняемого сценария, проверяющего текст по нескольким критериям, разработанного ранее и описанного также в предыдущей работе. В результате экспериментального исследования на выборке из 2500 статей, опубликованных в источниках ВАК/РИНЦ, были получены и математически обоснованы пороговые значения критериев. Было сформулировано необходимое, но не достаточное условие соответствия статьи научному стилю.

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 другое.

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 [3]. Вода - процентное соотношение стоп-слов и общего количества слов в тексте [3]. 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 [4], [5]. Соответствие распределения слов в тексте закону Ципфа говорит об уровне его естественности. Расчет 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 [6].

Более подробный обзор предметной области и вычисления критериев приведен в предыдущей статье [7].

Проблема

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

- 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 [4], [5] – λ .

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

Решение

Для получения данных для дальнейшего анализа был запущен исполняемый сценарий, разработанный ранее и описанных в предыдущей работе [7], на более крупной выборке научных статей. Анализ полученных значений числовых критериев позволит определить оправданные критерии оценки научных работ по трем числовым показателям.

Было проведено исследование на выборке из 2500 статей опубликованных в ВАК [8] и/или РИНЦ [9]. 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 [10], который выполняет веб-скрэпинг [11] научной интернет-библиотеки "Киберленинка" [12]. 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

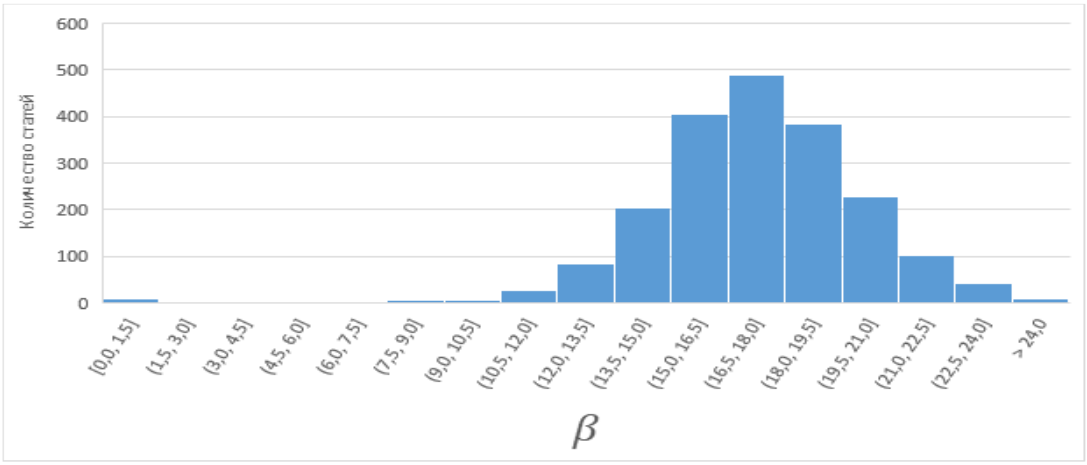

Рисунок 2 – Гистограмма распределения значений уровня водности текста статей из выборки

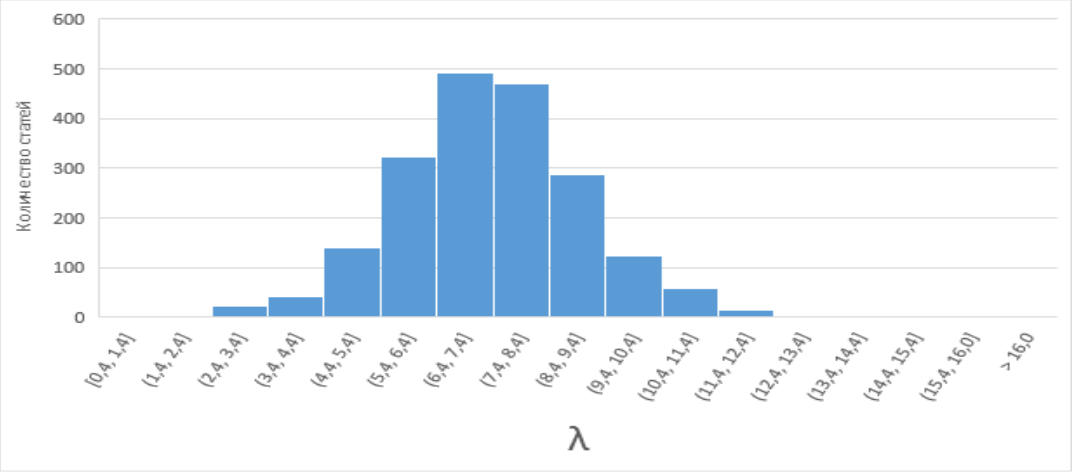

Рисунок 3 – Гистограмма распределения значений отклонения от идеальной кривой по Ципфу текста статей из выборки

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 [13], критерий Колмогорова-Смирнова [14], критерий Андерсона-Дарлинга [15]. В каждом из тестов проверяется нулевая гипотеза [16], о том, что каждая выборка получена из нормального распределения. Так, нулевая гипотеза считается верной до того момента, пока нельзя доказать обратное. Статистическая значимость [16] для тестов равна 0,05. P-значение [17] — величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы.

Использовалась [18] реализация тестов из статистической библиотеки SciPy [19]. На выходе каждый тест выдает два значения – D (статистика критерия для эмпирической функции распределения [14]) 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-Смирнов	3.09e-1	0.0
Андерсон-Дарлинг	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-Смирнов	2.29e-1	0.0
Андерсон-Дарлинг	1.4957e+1	7.866e-1

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

Критерий	D	P-значение
Шапиро	8.64e-1	3.512e-42
Колмогоров-Смирнов	1.29e-1	0.0
Андерсон-Дарлинг	2.8732e+1	7.866e-1

Как видно из результатов тестов – нет поводов отклонить нулевую гипотезу для каждой выборки, то есть 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 [20], 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. Установленные интервалы представлены в табл. 5.

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 корреляции. Был использован линейный коэффициент корреляции (коэффициент корреляции Пирсона) для расчета корреляции числовых критериев на основе полученных выборок:

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

где X и Y – значения критериев статьи, σ – среднеквадратичное отклонение, cov_{XY} - ковариация X и Y, $\bar{X}$ и $\bar{Y}$ - средние значения выборок.

Полученная матрица корреляции (матрица **не** нормирована):

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

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, в связи с общим знаменателем, которой можно пренебречь.

Запуски на тестовой выборке и других текстах

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 [21].

Принятые условия проверки оценок работ с помощью анализа критериев представлены в табл. 6.

Таблица 6 – Условия проверки оценок работ

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.

Для дальнейшего использования данных числовых критериев при оценке статьи, на базе полученных данных сформулируем необходимое условие соответствия научному стилю:

$$\alpha \in [6;14] \wedge \beta \in [14,20] \wedge \lambda \in [5.5,\ 9.5],$$

то есть все три числовых критерия должны попадать в установленные интервалы. Данное условие нужно считать необходимым, но не достаточным, в связи отсутствием анализа полезности содержания статьи.

Дополнительно было проведено исследование на текстах других жанров:

- работа «Корчеватель» [22], [23] – сгенерированная в научном стиле, не имеющая смысла статья, которая была принята для публикации в различные научные издания;
- популярные статьи в it-сообществе Хабр: «Моё разочарование в софте» [24], «Наши с вами персональные данные ничего не стоят» [25], «Рассказ о том, как я ворую номера

кредиток и пароли у посетителей ваших сайтов» [26], «Трехмерный движок на формулах Excel для чайников» [27];

- первый том «Капитала» Карла Маркса;
- роман «Идиот» Фёдора Достоевского;
- роман-поэма «Мёртвые души» Николая Гоголя;
- роман «Путешествие к центру Земли» Жюля Верна.

Результаты оценки представлены в табл. 7.

Таблица 7 – Результаты оценки текстов

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
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

По результатам проверки, значения всех трёх критериев статьи «Корчеватель» попали в установленные интервалы, т.е. работу можно считать соответствующей научному стилю.

Интернет-статьи и литературные произведения не написаны в научном стиле, и выделяются повышенным значением β.

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. Был сделан инструмент в виде исполняемого сценария, рассчитывающего данные критерии для статьи.

В дальнейшем планируется разработать веб-сервис, выполняющий проверку статей по полученным критериям, а также анализирующий форматирование статей и семантику на соответствие научному стилю [1].

Список литературы

1. Демидова А. К. Пособие по русскому языку: научный стиль, оформление научной работы. – Рус. яз., 1991.
2. Davis H. Search engine optimization. – " O'Reilly Media, Inc.", 2006.
3. Словарь терминов семантического анализа. // URL: seopult.ru/library (дата обращения 20.12.2018).
4. Newman M. E. J. Power laws, Pareto distributions and Zipf's law //Contemporary physics. – 2005. – Т. 46. – №. 5. – С. 323-351.
5. Lelu A. Jean-Baptiste Estoup and the origins of Zipf's law: a stenographer with a scientific mind (1868-1950) //Boletín de Estadística e Investigación Operativa. – 2014. – Т. 30. – №. 1. – С. 66-77.
6. Dong X. L. et al. Knowledge-based trust: Estimating the trustworthiness of web sources //Proceedings of the VLDB Endowment. – 2015. – Т. 8. – №. 9. – С. 938-949.
7. 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;
8. ВЫСШАЯ АТТЕСТАЦИОННАЯ КОМИССИЯ (ВАК) при Министерстве образования и науки Российской Федерации. // URL: vak.ed.gov.ru (дата обращения 20.12.2018).
9. РОССИЙСКИЙ ИНДЕКС НАУЧНОГО ЦИТИРОВАНИЯ. // URL: elibrary.ru/project_risc.asp (дата обращения 20.12.2018).
10. Исполняемый сценарий, получающий выборку статей. // URL: github.com/Eduard-Blees/Master-s-thesis/blob/master/script/leninka_scrapper.py (дата обращения 20.12.2018).
11. Boeing G., Waddell P. New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings //Journal of Planning Education and Research. – 2017. – Т. 37. – №. 4. – С. 457-476.
12. КиберЛенинка. Научная электронная библиотека, построенная на парадигме открытой науки // URL: cyberleninka.ru (дата обращения 20.12.2018).
13. Shapiro S. S., Wilk M. B. An analysis of variance test for normality (complete samples) //Biometrika. – 1965. – Т. 52. – №. 3/4. – С. 591-611.
14. Kolmogorov A. Sulla determinazione empirica di una lgge di distribuzione //Inst. Ital. Attuari, Giorn. – 1933. – Т. 4. – С. 83-91.
15. Anderson T. W., Darling D. A. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes //The annals of mathematical statistics. – 1952. – С. 193-212.
16. Гмурман Б. Е. Теория вероятностей и математическая статистика. – Москва «Высшая школа», 2003. – Т.478
17. Cumming G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better //Perspectives on Psychological Science. – 2008. – Т. 3. – №. 4. – С. 286-300.
18. Исполняемый сценарий, рассчитывающий математические критерии распределений. // URL: github.com/EduardBlees/Master-s-thesis/blob/master/script/results/testDistribution.py (дата обращения 20.12.2018).

19. SciPy module for Python // URL: scipy.org (дата обращения 20.12.2018).
20. Wheeler D. J. et al. Understanding statistical process control. – 1992. – T.406
21. Easton V. J., McColl J. H. Statistics glossary. – 2002.
22. Жуков М. С. Корчеватель: алгоритм типичной унификации точек доступа и избыточности. – 2008.
23. Stribling J., Aguayo D., Krohn M. Rooter: A methodology for the typical unification of access points and redundancy //Journal of Irreproducible Results. – 2005. – Т. 49. – №. 3. – С. 5.
24. Хабр. «Моё разочарование в софте» // URL: habr.com/post/423889/ (дата обращения 20.12.2018).
25. Хабр. «Наши с вами персональные данные ничего не стоят» // URL: habr.com/post/423889/ (дата обращения 20.12.2018).
26. Хабр. «Рассказ о том, как я ворую номера кредиток и пароли у посетителей ваших сайтов» // URL: habr.com/post/423889/ (дата обращения 20.12.2018).
27. Хабр. «Трехмерный движок на формулах Excel для чайников» // URL: habr.com/post/423889/ (дата обращения 20.12.2018).