**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра математического обеспечения и применения ЭВМ**

отчет

**по научно-исследовательской работе**

Тема: Система автоматической проверки наиболее частых формальных ошибок в научных текстах

|  |  |  |
| --- | --- | --- |
| Студент гр. 3304 |  | Блеес Э.И. |
| Руководители |  | Заславский М.М. |
|  |  | Кринкин К.В. |

Санкт-Петербург

2018

**ЗАДАНИЕ**

**на научно-исследовательскую работу**

|  |  |  |
| --- | --- | --- |
| Студент Блеес Э.И. | | |
| Группа 3304 | | |
| Тема НИР: Система автоматической проверки наиболее частых формальных ошибок в научных текстах | | |
| Задание на НИР:   * изучение возможности автоматизации параметров проверки научных статей на соответствие научному стилю; * разработка исполняемого сценария для проверки научной статьи по автоматизируемым критериям; * исследование разработанного сценария на опубликованных научных статьях для получения рекомендованных границ для критериев; | | |
| Сроки выполнения НИР: 13.02.2018 – 27.05.2018 | | |
| Дата сдачи отчета: 25.05.2018 | | |
| Дата защиты отчета: 25.05.2018 | | |
|  | | |
| Студент |  | Блеес Э.И. |
| Руководители |  | Заславский М.М. |
|  |  | Кринкин К.В. |

**Аннотация**

В данной статье приведено исследование возможности автоматизации процесса проверки научных статей на соответствие научному стилю, в результате которого было показано, что часть критериев проверки может быть автоматизирована. Было предложено решение по автоматизации процесса проверки научных статей в виде исполняемого сценария, проверяющего текст по нескольким критериям. Было проведено исследование решения на выборке статей, опубликованных в научных журналах ВАК и базе RSCI, в результате которого выявлены границы критериев автоматической проверки статей, которые будут использоваться в дальнейшем для рекомендаций при проверке.

**Summary**

This article explores the possibilities of automating the process of checking scientific articles for compliance with the scientific style, as a result of which it was shown that some of the verification criteria can be automated. A solution was proposed to automate the process of checking articles in the form of an executable script that verifies the text by several criteria. A study was conducted on a sample of articles published in the scientific journals of the Higher Attestation Commission and the RSCI database, as a result of which the boundaries of the verification criteria were identified, which will be used subsequently for recommendations in the audit of papers.

**содержание**

|  |  |  |
| --- | --- | --- |
|  | Введение | 5 |
| 1. | Обзор предметной области | 6 |
| 1.1. | Критерии сравнения аналогов | 7 |
| 1.2. | Выбор метода решения | 8 |
| 2. | Исполняемый сценарий | 10 |
| 2.1. | Описание метода решения | 10 |
| 2.2. | Сценарии использования | 10 |
| 2.3. | Алгоритм работы и используемые технологии | 10 |
| 2.4. | Исследование решения | 11 |
| 2.5. | Выводы | 13 |
|  | Заключение | 14 |
|  | Список использованных источников | 15 |
|  | Приложение А. Сведения о сборнике конференции | 16 |

**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 – 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 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.

**1. ОБзор преметной 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 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. 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. 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].

Рассмотрим возможность автоматизации критериев проверки текста. Уровень "воды" в тексте, его "тошнотность" и подсчет других числовых показателей, очевидно, можно автоматизировать. Но 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 [4].

Эти критерии можно применить и при проверке научных статей, но существуют веб-сервисы, проверяющие текст по этим критериям - сервисы, позволяющие провести SEO-анализ текста.

### 1y.ru

Анализатор качества контента [5]. Анализ проводится на базе закона Ципфа, то есть качество текста в данном случае определяется на основании соответствия частоты употребления слов в естественной речи и тексте. Результат выдается в двух окнах: в одном — график, в другом — частота использования отдельных слов и рекомендации по корректировке.

### text.ru

Сервис проверки текстов по многим параметрам, включая уникальность, проверку орфографии, выделение ключевых слов [6].

### contentmonster.ru

Сервис, осуществляющий поиск стоп-слов и подсчет их процентного соотношения к общей длине текста [7]. Определяет стоп-слова как всё то, что не несет самостоятельной смысловой нагрузки, но без чего не бывает связных текстов: предлоги, частицы, междометия, причастия, союзы, а также некоторые наречия, существительные и глаголы. Слишком большое количество таких слов затрудняет восприятие текста и увеличивает его водность.

## Критерии сравнения аналогов

Сравнение аналогов будет проводиться по следующим критериям:

* Многокритериальная проверка - как много критериев проверки использует сервис;
* Ограничение длины текста - отсутствие ограничения длины текста, поступающего на проверку;
* Проверка научного стиля - проверка текста на соответствие научному стилю.

В табл.1 представлено сравнение аналогов.

Таблица 1 - Сравнение аналогов

|  |  |  |  |
| --- | --- | --- | --- |
| Аналог | Многокритериальная проверка | Ограничение длины текста | Проверка научного стиля |
| 1y.ru | - | - | - |
| text.ru | + | + | - |
| contentmonster.ru | + | - | - |

Результаты сравнения показывают, что часть существующих сервисов предлагает многокритериальную проверку текста, при этом, не ограничивая его по длине. Но все аналоги осуществляют SEO-проверку, ни один из них не реализует проверку статьи на соответствие научному стилю.

## Выбор метода решения

Результаты сравнения аналогов показывают, что существует множество сервисов для SEO-проверки текста, но нет инструментов для проверки текста или статьи на соответствие научному стилю.

В связи с этим задачей является реализация решения, позволяющего автоматизировать проверку научных статей на соответствие научному стилю по нескольким критериям.

Метод решения - разработка исполняемого сценария. Данный метод выбран в связи с:

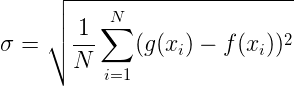
* Простотой разработки сценария;
* Легкостью поддержки решения;
* Легкостью запуска.

Реализуемые критерии проверки статьи:

* Анализ текста соответствию закону Ципфа [8-9] с рассчетом отклонения от идеального распределения;
* Проверка водности текста.

Закон Ципфа - эмпирическая закономерность распределения частоты слов естественного языка: 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.

Определение естественности текста осуществляется на основе отклонения графика частоты встречаемости слов от идеального графика по Ципфу. Закономерность Ципфа:  , где  - частота встречаемости n-го по рангу слова;  - количество повторений самого популярного слова в тексте. Показатель отклонения  вычисляется на основе среднеквадратичного отклонения точек дискретной функции  - практических показателей текста от точек дискретной функции  - идеальной функции по Ципфу. Показатель отклонения:



Данные критерии проверки были выбраны для реализации в первую очередь в связи с их наглядностью и простотой исправления замечаний автором проверяемой статьи. Решение является расширяемым и позволит в дальнейшем увеличить количество критериев проверки текста.

**2. исполняемый сценарий**

**2.1. Описание метода решений**

Исполняемый сценарий разработан на языке Python. Python выбран в связи с легкостью написания исполняемых сценариев на языке, а также наличием большого количества модулей для языка для разнообразных задач.

**2.2. Сценарий использования**

Предполагается единственный сценарий использования разработанного решения, состоящий из следующих этапов:

1. Запуск исполняемого сценария с указанием пути к директории с файлами для проверки;
2. Получение результата проверки.

В качестве выходных данных пользователь получает json файл с объектами, содержащими числовой показатель водности текста, числовой показатель тошноты текста и числовое значение отклонения графика частоты встречаемости слов в тексте от идеального графика по Ципфу. В дополнение к этому пользователю предоставляются рекомендации по интерпретации полученных результатов проверки

**2.3. Алгоритм работы и используемые технологии**

Поставленная задача требовала решения следующих подзадач:

* Получения текста из файла;
* Синтаксический разбор текста;
* Анализ текста как набора слов;
* Математические расчеты и построение графиков.

Получение текста из pdf, md, doc, docx файлов выполняется с помощью модуля textract. Полученный текст с помощью регулярного выражения разбивается на слова, получая список слов текста. Работа с регулярными выражениями осуществляется с помощью модуля re. Полученный список слов текста необходимо привести в нормальную языковую форму для дальнешей обработки, что возможно благодаря модулю pymorphy2 - морфологического анализатора для русского языка.

На данном этапе для определения "водности" текста необходимо подсчитать количество стоп-слов в нем, и исключить их для дальнейшей обработки. Список стоп-слов русского языка содержится в модуле nltk. Вычислительная работа с данными для их отображения осуществляется с помощью модуля scipy. Графики строятся средствами модуля matplotlib.

**2.4. Исследование решения**

Исследование решения проводилось на выборке из 1120 статей из журналов ВАК и базы RSCI, загруженных из ресурса Cyberleninka [10], с помощью вспомогательного скрипта. Так как эти статьи опубликованы в журналах ВАК и базе RSCI, можно опираться на результаты их проверки для получения рекомендованных интервалов критериев.

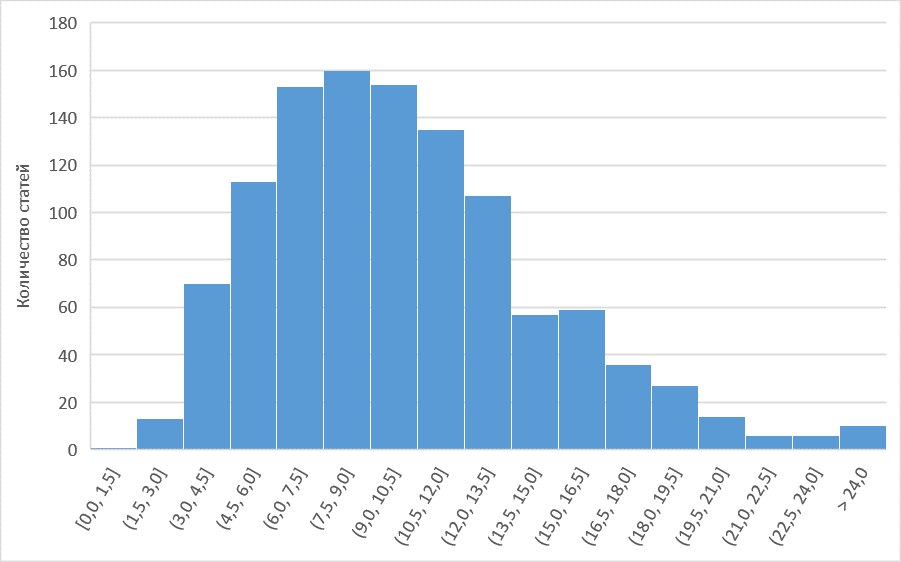


Рисунок 1 – Гистограмма числового показателя тошноты текста

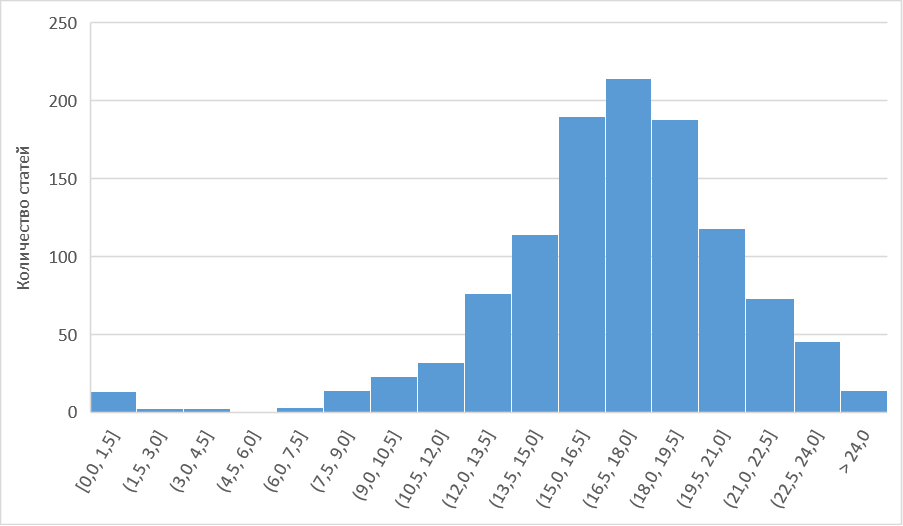


Рисунок 2 – Гистограмма числового показателя водности текста

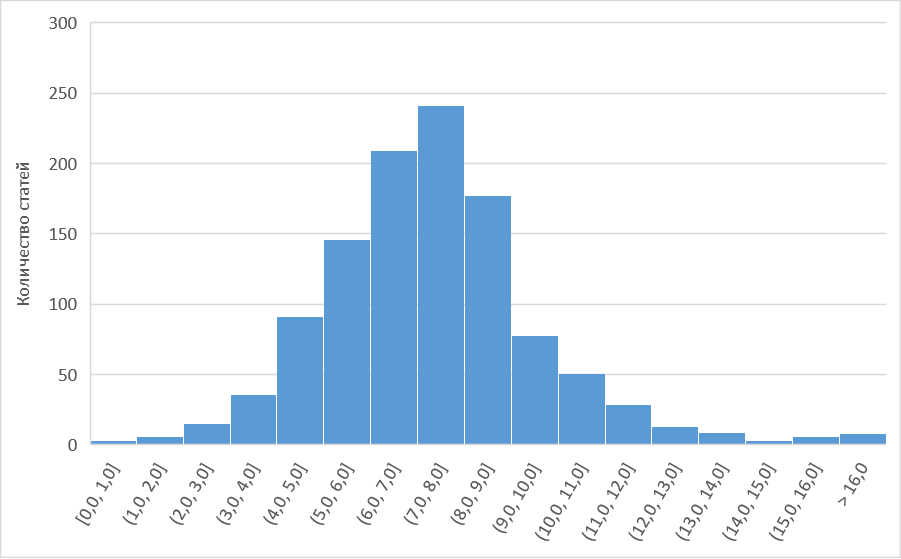


Рисунок 3 – Гистограмма числового показателя отклонения текста от закона Ципфа

Полученные данные свидетельствуют о наличии нормального распределения в выборке по каждому критерию. Рекомендуемый диапазон критерия – диапазон, в котором находится большая часть выборки, границы которого находятся вычетом и добавлением среднеквадратичного отклонения к медиане выборки. Полученные результаты представлены в табл. 2:

Таблица 2 - Полученные результаты

|  |  |  |  |
| --- | --- | --- | --- |
| Критерий | Медиана | Среднеквадратичное отклонение | Рекомендуемый диапазон |
| Тошнота текста | 9.42 | 4.53 | [ 4.89, 13.95] |
| Водность текста | 17.21 | 3.87 | [ 13.34, 21.08] |
| Отклонение от идеальной кривой по Ципфу | 7.18 | 2.66 | [ 4.52, 9.84] |

**2.4. Выводы**

В результате работы было проведено исследование возможности автоматизации процесса проверки научных статей на соответствие "научному стилю". Было предложено и реализовано решение в виде исполняемого сценария, позволяющее проверить научную статью на соответствие закону Ципфа, а также выполняющее расчет процентного соотношения стоп-слов к общему количеству слов в тексте. В качестве выходных данных пользователь получает числовой показатель "водности" текста, а также график соответствия текста закону Ципфа. В дополнение к этому пользователю предоставляются рекомендации по интерпретации полученных результатов проверки, полученные в результате проведения исследования решения на выборке статей, опубликованных в журналах ВАК и базе RSCI. В дальнейшем планируется увеличить количество критериев для проверки и реализовать веб-сервис.

**заключение**

В результате выполнения работы было проведено исследование решения на выборке статей, опубликованных в научных журналах ВАК и базе RSCI, в результате которого выявлены границы критериев автоматической проверки статей, которые будут использоваться в дальнейшем для рекомендаций при проверке.

В ходе выполнения НИР была написана статья «Автоматизация процесса проверки текста на соответствие научному стилю». Статья опубликована в сборнике конференции «Современные технологии в теории и практике программирования». Сведения о сборнике представлены в Приложении А.

**список использованных источников**

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
4. 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.
5. Сервис оценки качества текста. // URL: 1y.ru
6. Сервис оценки качества текста. // URL: text.ru
7. Сервис оценки качества текста. // URL: contentmonster.ru
8. Newman M. E. J. Power laws, Pareto distributions and Zipf's law //Contemporary physics. – 2005. – Т. 46. – №. 5. – С. 323-351.
9. 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.
10. Научная электронная библиотека «Киберленинка». // URL: cyberleninka.ru

**приложение А**

**Сведения о сборнике конференции**

