УДК 658.512.011.56:004.42

Э. И. Блеес (1 курс магистратуры, каф. МОЭВМ, СПбГЭТУ «ЛЭТИ»),

В.Ю. Андросов (4 курс бакалавриата, каф. МОЭВМ, СПбГЭТУ «ЛЭТИ»),

М. М. Заславский (ассистент каф. МОЭВМ, СПбГЭТУ «ЛЭТИ»)

АВТОМАТИЗАЦИЯ ПРОЦЕССА ПРОВЕРКИ ТЕКСТА НА СООТВЕТСТВИЕ НАУЧНОМУ СТИЛЮ

Проблема и её актуальность

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.

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 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. 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].

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, например Анализатор качества контента 1y.ru [5], сервис проверки текстов text.ru [6], сервис, осуществляющий поиск стоп-слов и подсчет их процентного соотношения к общей длине текста contentmonster.ru [7].

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	+	-	-

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 и легкостью поддержки решения.

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. Данные критерии проверки были выбраны для реализации в первую очередь в связи с их наглядностью и простотой исправления замечаний автором проверяемой статьи.

В связи с наличием в открытом доступе репозитория [10], в котором студенты СПбГЭТУ кафедры МОЭВМ пишут статьи в файлах формата .md, и необходимостью тестирования решения логично указывать исполняемому сценарию путь к директории в .md файлами, для получения из них текста и дальнейшего анализа. В итоге, необходимо разработать исполняемый сценарий, получающий на вход путь к директории, в которой находятся файлы, содержащие текст, и выводящий результат проверки.

Описание метода решения

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. В качестве выходных данных пользователь получает числовой показатель водности текста, а также график соответствия текста закону Ципфа и числовое значение отклонения от него графика частоты встречаемости слов в тексте. В дополнение к этому пользователю предоставляются рекомендации по интерпретации полученных результатов проверки. Пример графика представлен на рис. 1.

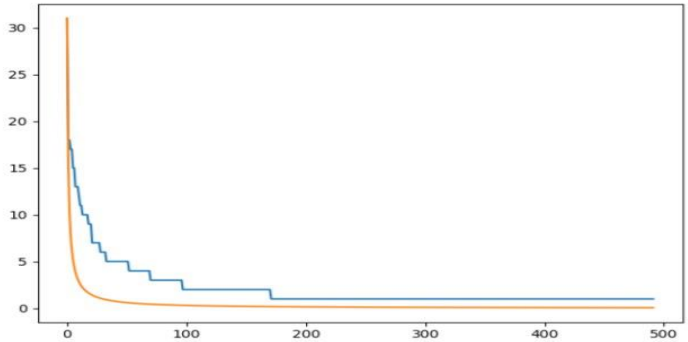

Рисунок 1 - Пример графика

Алгоритм работы и используемые технологии

Поставленная задача требовала решения следующих подзадач:

1. Синтаксический разбор .md файлов;
2. Синтаксический разбор текста;
3. Анализ текста как набора слов;
4. Математические расчеты и построение графиков.

При запуске исполняемого сценария находятся все .md файлы в директории, которые поступают на обработку, осуществляемую с помощью модуля mistune [11], переводящего .md файл в html документ. Это удобно, в связи с развитостью html-парсеров по причине огромной популярности и распространенности формата. Из html документа выделяется весь текст, который затем с помощью регулярного выражения разбивается на слова, получая список слов текста. Полученный список слов текста необходимо привести в нормальную языковую форму для дальнейшей обработки, что осуществляется с помощью pymorphy2 [13] - морфологического анализатора для русского языка. Список стоп-слов русского языка содержится в модуле nltk [14]. Вычислительная работа с данными для их отображения осуществляется с помощью модуля scipy [15]. Графики строятся средствами модуля matplotlib [16].

Исследование решения

Было проведено исследование работы исполняемого сценария на статьях, написанных студентами СПбГЭТУ "ЛЭТИ" в рамках факультатива по подготовке научных статей [10], а также для дальнейшего сопоставления результатов, в качестве примера опубликованной, а соответственно и прошедшей проверку, статьи, была взята работа Заславского М. М., Блееса Э. И., Баландина С. И. - "Метод обработки в реальном времени открытых данных, содержащих геоконтекстную разметку"[17].

Результаты исследования

В результате исследования были получены средние показатели рассчитываемых параметров, которые необходимо будет стандартизировать, проведя исследования на крупной выборке опубликованных статей. Была замечена тенденция к понижению уровня естественности в научно-технических статьях, которую можно объяснить частым упоминанием ключевых для понимания текста связок слов. Повышенный уровень водности объясняется необходимостью в связи частей статьи в единое целое, для лучшего понимания читателя.

Заключение

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. Было проведено исследование решения на статьях, написанных студентами СПбГЭТУ. В дальнейшем планируется увеличить количество критериев, в том числе реализовать проверку частоты употребления в тексте ключевых слов, а также провести исследования на крупной выборке опубликованных статей.

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. Репозиторий факультатива по подготовке научных статей. // URL: github.com/moevm/scientific_writing-2017
11. Mistune module for Python // URL: github.com/lepture/mistune
12. Bs4 module for Python // URL: crummy.com/software/BeautifulSoup/bs4/doc/
13. PyMorphy2 module for Python // URL: github.com/kmike/pymorphy2
14. Nltk module for Python // URL: nltk.org
15. SciPy module for Python // URL: scipy.org
16. MatPlotLib module for Python // URL: matplotlib.org
17. Заславский М.М., Блеес Э.И., Баландин С.И. Метод обработки в реальном времени открытых данных, содержащих геоконтекстную разметку // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 5. С. 850–858. doi: 10.17586/2226-1494-2017-17-5-850-858