Исследование возможности автоматизации процесса проверки текста на соответствие научному стилю

Аннотация

В данной статье приведены 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.

Введение

Соответствие статьи научному стилю является одним из основных критериев принятия статьи на публикацию. 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].

SEO (search engine optimization) анализ 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. [2]

Тошнота – это показатель повторений в текстовом документе ключевых слов и фраз. Синонимом тошноты является термин плотность [2].

Стоп-слова – это слова в тексте, которые не несут смысловой нагрузки. Иначе их называют также шумовые слова [2].

Вода - процентное соотношение стоп-слов и общего количества слов в тексте [2].

Эти критерии можно применить и при проверке научных статей.

Существуют веб-сервисы, позволяющие провести SEO-анализ текста.

1y.ru

Анализатор качества контента [3]. 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 [4].

contentmonster.ru

Сервис, осуществляющий поиск стоп-слов и подсчет их процентного соотношения к общей длине текста [5]. 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

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

Таблица сравнения по критериям

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

Выводы по итогам сравнения

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 - исполняемый сценарий. Данный метод выбран в связи с:

- Простотой разработки сценария;

- Легкостью поддержки решения;
- Легкостью запуска.

Реализуемые критерии проверки статьи:

- Анализ текста соответствию закону Ципфа;
- Проверка водности текста.

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

Необходимо разработать исполняемый сценарий, получающий на вход путь к директории, в которой находятся файлы, содержащие текст, и выводящий результат проверки.

Описание метода решения

Метод решения - исполняемый сценарий, написанный 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.

Входные данные

Аргументом командной строки при запуске исполняемого сценария указывается путь к директории, в которой находятся файлы, содержащие текст для проверки. Сценарий будет учитывать все файлы с расширением .md находящиеся в папке.

Сценарий использования

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

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.

Алгоритм работы и используемые технологии

Поставленная задача требовала решения следующих подзадач:

1. Парсинг .md файлов;
2. Парсинг 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, переводящего .md файл в html документ. Это удобно, в связи с развитостью html-парсеров по причине огромной популярности и распространенности формата. Так же идея перевода файла в промежуточный формат html позволит в дальнейшем добавить поддержку анализа текстов в другом формате. Парсинг html документа осуществляется с помощью модуля bs4.

Из html документа выделяется весь текст, который затем 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 неободимо 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.

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 так же выполняющее рассчет 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.

Поставленные цели были достигнуты.

В дальнейшем планируется увеличить количество критериев, в том числе реализовать проверку частоты употребления в тексте слов, составляющих семантическое ядро.

Список литературы

1. Демидова А. К. Пособие по русскому языку: научный стиль, оформление научной работы. – Рус. яз., 1991.
2. Словарь терминов семантического анализа. // URL: seopult.ru/library
3. Сервис оценки качества текста. // URL: 1y.ru
4. Сервис оценки качества текста. // URL: text.ru
5. Сервис оценки качества текста. // URL: contentmonster.ru