Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина)

Разработка системы автоматизированной проверки наиболее частых ошибок в научных текстах

Выполнил: Блеес Эдуард Игоревич, гр. 3304
Руководитель: Заславский Марк Маркович, ассистент

Актуальность

Процесс проверки статей изданиями в текущем виде:

- Долгая переписка с рецензентом и редакторами;
- Повторные отправки после малейших исправлений.

Существует курс по написанию научных статей на Stepik, для которого необходима частичная автоматизация проверки статей

Цель и задачи

Цель: Разработать программу для проверки статьи на соответствие научному стилю и поиску наиболее частых ошибок в ней.

Задачи:

- Исследовать возможность автоматизации проверки научных статей на соответствие научному стилю;
- Построить математическую модель проверки статьи;
- Провести экспериментальное исследование для определения допустимых значений критериев;
- Реализовать программный прототип решения.

Исследование возможности автоматизации проверки статей на соответствие научному стилю

В результате обзора были выделены морфологические особенности научного стиля:

- Использование личных местоимений. Личные и притяжательные местоимения (я, ты, мною, вы, наш) имеют отвлеченно-обобщенный характер и их употребление необходимо избегать;
- Использование неопределенных местоимений (кое-что, что-нибудь). Эти местоимения, в силу их неопределенности, не употребляются.

Исследование возможности автоматизации проверки статей на соответствие научному стилю

Проверка качества текста или соответствие информационному стилю. SEO-анализ.

Вводимые термины:

- Тошнота – это показатель повторений в текстовом документе ключевых слов и фраз. Синонимом тошноты является термин плотность.
- Стоп-слова – это слова в тексте, которые не несут смысловой нагрузки
- Вода - процентное соотношение стоп-слов и общего количества слов в тексте
- Эмпирическая закономерность распределения частоты слов естественного языка - Закон Ципфа

Исследование возможности автоматизации проверки статей на соответствие научному стилю

Информационный стиль и SEO-анализ вводят морфологические ограничения:

- Использование слов усилителей (безусловно, очень, абсолютно и др.);
- Использование обобщений (со всего мира, весь, в общем);
- Необъективная оценка (уникальный, новейший);
- Использование риторических вопросов.

Исследование возможности автоматизации проверки статей на соответствие научному стилю

Автоматизируемые правила проверки научных статей в существующем курсе:

- Каждое ключевое слово упоминается в основном тексте хотя бы один раз;
- Более половины элементов списка литературы - актуальные и значимые научные работы;
- Все элементы списка литературы имеют минимум одно упоминание в тексте;
- Все рисунки и таблицы имеют подрисуночные подписи и ссылки в тексте.

Исследование возможности автоматизации проверки статей на соответствие научному стилю

Обзор аналогов

Аналог	Многокритериальная проверка	Нет ограничения на длину текста	Проверка стиля	Возможность загрузки файлов для проверки
1y.ru	-	+	-	-
text.ru	+	-	-	-
contentmonster.ru	+	+	-	-
glvrd.ru	+	+	+	-

Экспериментальное исследование

Выборка из 2500 статей опубликованных в источниках ВАК или РИНЦ.

Проверяемая гипотеза:
Качество научной статьи влияет на значения определенных числовых критериев, а также полученная выборка значений критериев соответствует нормальному распределению

Результаты экспериментального исследования

Числовые критерии:

- Тошнота текста – α;
- Уровень воды в тексте – β;
- Значение отклонения текста статьи от идеальной кривой по Ципфу – λ.

Экспериментально установленные интервалы:

Критерий	Интервал
α	[6, 14]
β	[14, 20]
λ	[5.5, 9.5]

Математическая модель проверки статьи

В результате исследования было выделено:

- 3 рассчитываемых числовых критерия;
- 5 типов проверяемых стилистических ошибок;
- 6 типов проверяемых структурных ошибок.

Оценка статьи:

$$К = В - Ф$$

Где К – оценка статьи, В – базовое значение К, Ф – штраф.

Математическая модель проверки статьи

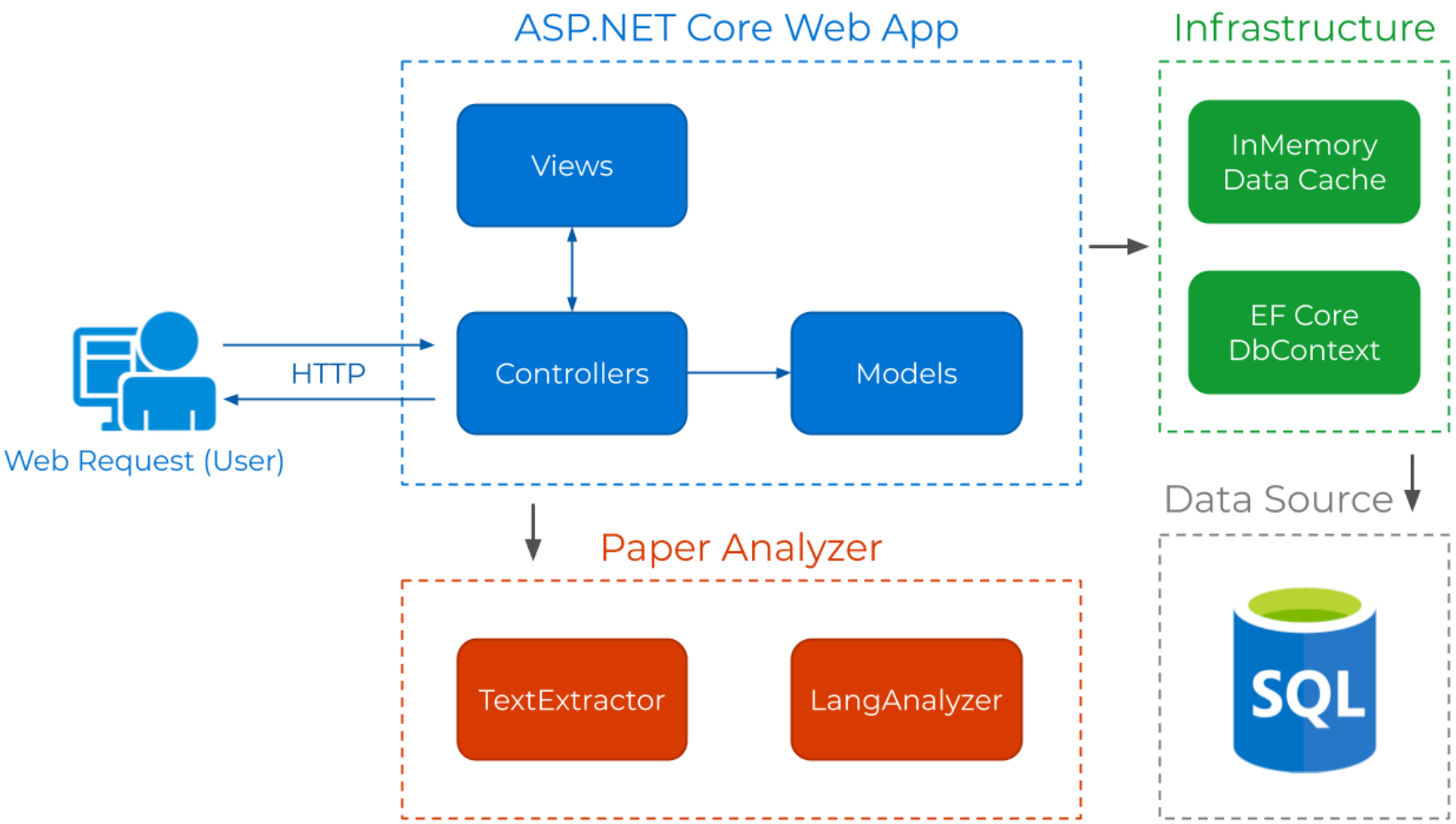
$$\mathrm{E}(\alpha) = \begin{cases} 1, \alpha \in [6;14] \\ 0, \alpha \notin [6;14] \end{cases}$$

$$\mathrm{B} = C_1 \times \mathrm{E}(\alpha) + C_2 \times \mathrm{E}(\beta) + C_3 \times \mathrm{E}(\lambda)$$

$$\Phi = C_4 \times N_1 + C_5 \times N_2$$

Где E – попадание критерия в установленный промежуток, N_1 – количество структурных ошибок, N_2 – количество стилистических ошибок, C_1 ,C_1, C_2, C_3, C_4, C_5 – коэффициенты.

Разработанное решение



Заключение

- Было проведено исследование возможности автоматизации проверки научных статей на соответствие научному стилю, по результатам которого были выделены критерии проверки статей;
- На основании проведенного обзора была построена математическая модель проверки статьи, включающая в себя проверку числовых критериев, и поиск структурных и стилистических ошибок;
- Было проведено экспериментальное исследование на статьях, опубликованных в источниках ВАК или РИНЦ, по результатам которого были определены допустимые значения критериев и была настроена и формализована модель;
- Было проведено экспериментальное исследование на статьях и произведениях других жанров для проверки корректности полученной модели, показавшее корректность разработанной модели проверки;
- Было разработано решение в виде веб-сервиса.

Апробация работы

- 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 - 2018. - С. 118-121;
- Блеес Э.И., Заславский М.М. Исследование критериев соответствия текста научному стилю // Научно-технический вестник информационных технологий, механики и оптики. 2019. Т. 19. № 2. С. 299–305. doi: 10.17586/2226-1494-2019-19-2-299-305;
- Репозиторий проекта https://github.com/EduardBlees/Master-s-thesis.

Дополнительный слайд №1. Собственно-научный подстиль

В рамках данной работы была реализована проверка статей на соответствие собственно-научному подстилю. Собственно-научный подстиль — академическое изложение, адресованное специалистам.

Характеристики:

- Точность передаваемой информации;
- Убедительность аргументации;
- Логическая последовательность изложения;
- Лаконичность.

Цель подстиля — выявление и описание новых фактов, закономерностей, открытий.