Санкт-Петербургский государственный университет

Группа 20.Б08-мм

Нафикова Лиана Ирековна

Расширение функциональности валидатора текстов выпускных квалификационных работ

Отчёт по учебной практике
в форме «Производственное задание»

Научный руководитель:
ассистент Чернышев Г.А.

Санкт-Петербург
2023

Оглавление

Введение

Одним из итогов обучения студента в вузе является выпускная квалификационная работа. Также во время учебного процесса приходится сталкиваться с написанием курсовых, эссе, рефератов. Их проверка — очень трудозатратный процесс как для студентов, так и для научных руководителей. Нужно следить не только за качеством содержательной части, но и за следованием стандартам оформления. Последнее частично поддаётся автоматизации, которая обеспечивает более сильный контроль соблюдения формальных норм проверки ВКР.

Mundane Assignment Police — веб-приложение, реализованное с на языке программирования Kotlin, помогающее в обработке ВКР, указывая на самые распространённые ошибки в загруженных PDF-файлах. Оно предназначено для обработки работ студентов Математико-механического факультета СПбГУ. В данный момент находится в стадии разработки.

В приложении создаются правила проверки текста. Расширение приложения происходит за счёт увеличения количества добавляемых правил.Таким образом, в этой работе предлагается улучшение работы Mundane Assignment Police, с помощью которого будет возможно проверять выпускные квалификационные работы на наличие ошибок оформления.

1. Постановка задачи

Целью работы является расширение функционала веб-приложения для проверки PDF-файлов с текстами выпускных квалификационных работ.

Для её выполнения были поставлены следующие задачи:

1. организовать сбор статистики, помогающей проводить анализ работы;
2. реализовать алгоритм, проверяющий на равенство количества пунктов в постановке задач и результатов в заключении;
3. провести анализ результатов выполненных алгоритмов на базе имеющихся работ.

2. Обзор

В данном разделе рассматриваются работы и методы, решающие проблему проверки и анализа текста на соответствие выдвинутым стандартам.

2.1. Методы обработки текста

Необходимость обработки текста послужила причиной появления большого количества работ, проверяющих соответствие текста определённым правилам, такими как, например, проверка орфографии.

В работе [5] используют статистический анализ (Statistical Analysis), Stripping, полный поиск (Complete Lookup). Первый метод основан на частотном анализе соседних символов в словах, которые необходимо проверить. Второй состоит из удаления префиксов и суффиксов до тех пор, пока слово не будет сокращено до корня, который затем проверяется в словаре. Третий же заключается в наличии большого словаря, который содержит все формы слов, которые следует рассматривать.

Также существуют всевозможные синтаксические подходы к обработке текста. В [4] используются синтаксический анализ (без учёта семантики слов), словарная информация (значение слова и его связь с другими), создание базы знаний, т.е. абстрактное представление тематической области или конкретной среды, включая основные понятия, представляющие интерес в этой области, и различные отношения между сущностями.

2.2. Аналоги

Прямой аналог разрабатываемого приложения не был найден. Тем не менее, в разных предметных областях существуют работы, осуществляющие различные проверки соответствия стандартам или определённым значениям.

Похожие системы встречаются в юридической сфере. Так, например, в Formalex [3], рассматривается конкретная структура юридиче-

ских документов. Авторы анализируют тексты законов и на их основе строят логические принципы, проверяют их непротиворечивость.

Существует также система CLAWS [1] (автоматическая система тегирования слов с постоянным правдоподобием) разработанная для грамматического анализа. CLAWS можно модифицировать обнаруживать грамматические ошибки. Полученная система не может обнаружить все ошибки в типизированных документах; но также и гораздо более сложные системы, которые пытаются выполнить полный анализ, требуют гораздо больших вычислений.

В [2] же рассматриваются алгоритмы, которые помогают определить, правильно ли используются ли ссылки в Википедии[1]. Чтобы решить проблему ошибочных ссылок, авторы изучили семантические отношения между текстами и сущностями в Википедии. На основе множества эвристик вводится специальная метрика, по которой определяется, является ли ссылка правильной или нет. Задача разделена на два этапа: генерация потенциальных ссылок на ошибки и классификация и исправление ссылок.

[1]https://en.wikipedia.org/wiki/Wikipedia — Википедия (дата обращения 06.06.2022)

3. Метод

В этом разделе описывается процесс сбора статистики и реализации алгоритма, обнаруживающего несовпадение пунктов в постановке задач и заключении.

3.1. Сбор статистики

Статистика включает в себя:

- Анализ слов текста ВКР;
- Анализ размера секций работы.

Чтобы наиболее точно посчитать слова в тексте, было необходимо определить, какие наборы символов, нужно считать словами, а какие нет. Список слов:

- наборы русских или английских символов, у которых однозначно определяется лемма;
- слова-сокращения;
- слова, которые написаны с переносом строки;
- названия (имена, языки программирования, названия книг и др.);
- прямые ссылки на другие источники.

Наборы символов, которые не считались словами при проверке:

- номера страниц, числа;
- предлоги, частицы и другие стоп-слова[2];
- знаки препинания.

Для реализации этого алгоритма было необходимо провести:

[2]https://ranks.nl/stopwords — словарь стоп-слов, используемых в работе (дата обращения: 30.03.2022).

1. Поиск слова в словаре стоп-слов;
2. Поиск слова в словаре ключевых слов;
3. Склеивание частей слова, которое разбилось на две части в результате переноса со строки на строку;
4. Удаление знаков препинания и спец.символов в словах;
5. Удаление цифр.

Чтобы это сделать, были осуществлены проверки, основанные на анализе символов, входящих в строку, содержащую проверяемое слово (для пунктов 3–5), также проведён анализ списков (для 1–2).

Статистика секций представляет собой подсчёт размера каждой секции в страницах и процентное отношение размера секции к размеру всей работы без титульного листа. Секции находились на основе алгоритмов, уже реализованных в проекте.

3.2. Проверка равенства количеств пунктов в постановке задач и результатов в заключении

В первую очередь, понадобилось расширить функционал правил, работающих со списками. Были добавлены возможности отфильтровывать проверяемые списки определённым образом (в частности, для конкретного правила списки фильтровались по тому, на каких страницах они находятся), работа с несколькими списками.

Был реализован поиск секции с задачами в оглавлении. В случае, если в названиях секций не обнаружены слова «задачи», «задач» или «Задачи» (они находятся либо в ведении, либо не поставлены), то выдаётся ошибка «Задачи не выделены в содержании».

Также понадобилось организовать поиск страниц, на которых находятся задачи и заключение, чтобы однозначно определить, какие списки нужно сравнивать. Это также делалось с помощью нахождения нужных секций в документе.

Если в одной из секций не найдены списки, то подчёркиваются соответствующие названия секций. Если в оба списка пусты, подчеркиваются названия обеих секций. В случае, если и задачи, и результаты обнаружены, то списки из заключения фильтруются по количеству пунктов в задачах. Если списков, больших или равных по размеру списку задач не обнаружено, то считается, что подходящих списков нет. Также выдаётся ошибка «Задачи и результаты не совпадают», и подчёркивается первый неподходящий список из заключения.

4. Эксперимент

Тестирование проводилось на базе данных курсовых, магистерских и бакалаврских работ[3] Математико-механического факультета СПбГУ. Из них приложение обработало 19 работ со следующими результатами (таблица 1, таблица 2, таблица 3).

4.1. Статистика

Таблица 1. Статистика размеров секций

Тип работы	Размер	Введение	Обзор	Заключение
Курсовые	32	2 (6,25%)	8 (25,00%)	1 (3,12%)
	28	2 (7,14%)	1 (3,57%)	4 (14,29%)
	14	2 (14,28%)	4 (28,57%)	2 (14,28%)
	14	1 (7,15%)	4 (28,57%)	2 (14,28%)
	27	2 (7,41%)	4 (14,81%)	4 (14,81%)
	21	2 (9,52%)	5 (23,81%)	2 (9,52%)
	17	1 (5,88%)	1 (5,88%)	2 (11,76%)
	13	1 (7,69%)	2 (15,38%)	3 (23,08%)
Бакалаврские	38	3 (7,89%)	10 (26,31%)	6 (15,79%)
	19	1 (5,26%)	2 (10,53%)	2 (10,53%)
	49	2 (4,08%)	2 (4,08%)	9 (18,37%)
	31	2 (6,45%)	7 (22,58%)	3 (9,68%)
	26	1 (3,85%)	7 (26,92%)	3 (11,54%)
	32	3 (9,38%)	4 (12,50%)	4 (12,50%)
Магистерские	48	2 (4,16%)	15 (31,25%)	1 (2,08%)
	41	3 (7,32%)	10 (24,39%)	4 (9,75%)
	37	3 (8,10%)	7 (18,91%)	7 (18,91%)
	33	3 (9,09%)	9 (27,27%)	4 (12,12%)
	47	3 (6,38%)	6 (12,77%)	3 (6,38%)

На основе полученной статистики также можно выдавать предупреждения в случае, если размер секций больше/меньше рекомендуемых.

[3]https://github.com/Darderion/map-dataset — репозиторий с базой выпускных квалификационных работ на Github (дата обращения: 15.04.2022).

Таблица 2. Статистика по количеству слов в ВКР			
Тип работы	Всего	Уникальные слова	Самое встречаемое
Курсовые	4270	1539	Образцом (75)
	1951	778	Пакета (31)
	957	483	Продолжение (17)
	878	463	Сигнала (33)
	1784	787	Поиска (52)
	1373	557	Протоколов (46)
	1225	529	Гибернации (50)
	688	387	Конфигурации (19)
Бакалаврские	2605	1065	Алгоритмов (69)
	1257	535	Поворота (37)
	4712	1651	Сообщений (85)
	2453	981	Алгоритма (78)
	1807	750	Устройств (52)
	2227	908	Вещей (49)
Магистерские	3125	1320	Платформа (49)
	3649	1392	Сети (81)
	3148	1450	Структуры (38)
	2354	975	Редактора (44)
	3921	1264	Состояний (87)

Статистика показывает, что стоп-слова увеличивают работу в среднем в два-три раза. При этом, статистика не является абсолютно точной, т.к. из строк слов при её составлении удаляются некоторые символы и первая буква переводится в верхний регистр. Т.е. например, ссылки будут отображены некорректно.

4.2. Совпадение задач и результатов

Таблица 3. Задачи и результаты

Тип работы	Ожидаемый результат	Реальный результат	Особенности работы
Курсовые	Задачи не выделены	Совпадает	
	Нет ошибок	Совпадает	
	Задачи и результаты не совпадают	Совпадает	Нет списков в заключении
	Нет ошибок	Совпадает	
	Задачи не выделены	Задачи и результаты не совпадают	В оглавлении секция, содержащая слово «задач» В секции нет списков
	Нет ошибок	Совпадает	
	Нет ошибок	Совпадает	
Бакалаврские	Задачи и результаты не совпадают	Совпадает	Результатов меньше
	Нет ошибок	Совпадает	
	Задачи не выделены	Совпадает	
	Нет ошибок	Совпадает	
	Нет ошибок	Совпадает	
	Нет ошибок	Совпадает	
Магистерские	Нет ошибок	Совпадает	
	Нет ошибок	Совпадает	Два списка в заключении
	Нет ошибок	Совпадает	Два списка в заключении
	Нет ошибок	Совпадает	
	Задачи не выделены	Совпадает	

Работы были просмотрены на наличие ошибок. В втором столбце описан ожидаемый результат проверки. В случае если реальный реазультат совпадает с ожидаемым, в третьем столбце пишется «Совпадает», иначе правило, которое сработало. В третьем столбце указаны особенности конкретной работы.

Заключение

В ходе работы были выполнены следующие задачи:

1. организация сбора статистики,помогающей проводить анализ работы;

2. реализация алгоритма, проверяющего на равенство количества пунктов в постановке задач и результатов в заключении;

3. проведён анализ результатов выполненных алгоритмов на базе имеющихся работ.

Открытый код проекта доступен по ссылке на репозиторий[4] Github. В дальнейшем, в проект будет добавлена проверка других правил.

[4]https://github.com/Liana2707/map — репозиторий проекта на Github (дата обращения: 22.03.2022).

Список литературы

[1] Atwell Eric Steven. How to Detect Grammatical Errors in a Text without Parsing It // Proceedings of the Third Conference on European Chapter of the Association for Computational Linguistics. — EACL '87. — USA : Association for Computational Linguistics, 1987. — P. 38–45. — URL: https://doi.org/10.3115/976858.976865.

[2] Error Link Detection and Correction in Wikipedia / Chengyu Wang, Rong Zhang, Xiaofeng He, Aoying Zhou // Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. — CIKM '16. — New York, NY, USA : Association for Computing Machinery, 2016. — P. 307–316. — URL: https://doi.org/10.1145/2983323.2983705.

[3] Performance Improvement on Legal Model Checking / Carlos Faciano, Sergio Mera, Fernando Schapachnik et al. // Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law. — ICAIL '17. — New York, NY, USA : Association for Computing Machinery, 2017. — P. 59–68. — URL: https://doi.org/10.1145/3086512.3086518.

[4] Salton G., Smith M. On the Application of Syntactic Methodologies in Automatic Text Analysis // SIGIR Forum. — 1989. — may. — Vol. 23, no. SI. — P. 137–150. — URL: https://doi.org/10.1145/75335.75479.

[5] Turba Thomas N. Checking for Spelling and Typographical Errors in Computer-Based Text // Proceedings of the ACM SIGPLAN SIGOA Symposium on Text Manipulation. — New York, NY, USA : Association for Computing Machinery, 1981. — P. 51–60. — URL: https://doi.org/10.1145/800209.806454.