

Разработка онлайн веб-приложения для ведения заметок с автоматическим выделением ключевых слов и подбором рекомендаций о сотрудничестве пользователей

Логвинов Ю.А., Пономарев И.В.

*Алтайский государственный университет, г. Барнаул
slymerrow@gmail.com, igorpon@mail.ru*

Аннотация

Работа посвящена созданию онлайн веб-приложения для ведения заметок, позволяющему пользователю выполнить автоматическое извлечение ключевых слов из текстов его заметок и на основании данного набора получить рекомендации о сотрудничестве с другими клиентами сервиса.

Ключевые слова: разработка веб-приложения, разработка веб-сайта, разработка веб-сервиса, ведение заметок, выделение ключевых слов.

1. Введение

В настоящее время среди пользователей сети Интернет большой популярностью пользуются веб-приложения для ведения заметок. Они позволяют вести записи на самые разные темы – от распорядка дня до научных конспектов. Однако не каждый сервис для ведения личных записей позволяет пользователю внедрять в тексты заметок гиперссылки на другие заметки. Между тем, такая функция, на наш взгляд, может оказаться весьма полезной при ведении научных записей. Возможность быстро перейти от одной заметки к другой, кликнув по гиперссылке, позволяет видеть контекст каждой заметки и легче ориентироваться в своих записях.

Ещё одной полезной опцией, которую могло бы предоставить подобного рода приложение, является возможность автоматического выделения ключевых слов из текстов заметок. Впоследствии информация о ключевых словах и частотах их встречаемости может быть использована для определения ранговых систем ключевых слов и предоставления рекомендаций относительно сотрудничества с другими пользователями на основании числа общих ключевых слов и силы корреляции между их ранговыми системами. Таким образом можно будет выявить сходство интересов разных пользователей.

Среди существующих приложений, разработанных для схожих задач, можно выделить следующие:

- Evernote – веб-сервис и набор программного обеспечения для создания и хранения заметок. В качестве заметки может выступать фрагмент форматированного текста, веб-страница целиком, фотография, аудиофайл, рукописная запись или вложения с файлами другого типа. Приложение поддерживает создание гиперссылок между заметками. Система веб-сервиса сосредоточена на том, чтобы предлагать релевантную информацию из интернета, а не сотрудничество с другими пользователями. Ссылка на ресурс: <https://evernote.com/ru>.

- Notion – приложение, которое предоставляет такие компоненты, как базы данных, доски канбан, вики, календари и напоминания. Пользователи могут подключать эти компоненты для создания собственных систем управления знаниями, ведения заметок, управления данными, управления проектами и другими. Данное приложение предоставляет возможность совместного ведения этих компонентов группой пользователей. Ссылка на ресурс: <https://www.notion.so/>.
- Roam Research – веб-приложение для ведения заметок, в основе системы которого лежит направленный граф, что освобождает его от ограничений классической файловой системы. Ссылка на ресурс: <https://roamresearch.com/>.

2. Программная реализация

Для создания веб-приложения были выбраны следующие средства разработки:

- Язык разметки гипертекстовых документов HTML – базовый язык разметки в сфере веб-разработки, используемый для структурирования и отображения страниц сайтов.
- Каскадные таблицы стилей CSS – инструмент, позволяющий отделить содержание документа от его визуального оформления.
- Язык разработки JavaScript – динамический скриптовый язык программирования высокого уровня, поддерживаемый всеми современными браузерами. Язык легко интегрируется с вёрсткой (HTML) и сервером.
- Скриптовый язык серверного программирования PHP, обладающий хорошей совместимостью с различным операционными системами и серверами. Функционал приложений на PHP может быть сравнительно легко расширен благодаря подключению необходимых библиотек.
- Система управления базами данных PostgreSQL, среди преимуществ которой можно выделить открытый исходный код и совместимость с языком запросов SQL.
- В качестве среды разработки был выбран редактор исходного кода VSCode, поддерживающий множество языков программирования и предоставляющий программисту удобны интерфейс для работы с кодом.

Что касается структуры приложения, было принято решение хранить файлы на каждом языке, упомянутом выше, в соответствующем каталоге файловой системы проекта. Все изображения, используемые при создании сайта, хранятся в папке `/images/`. Папка `fonts` содержит используемые на сайте шрифты. Код подключения к базе данных также вынесен в отдельный файл `ZettelSQL.txt`, исходя из соображений дальнейшего развития приложения.

При проектировании базы данных веб-сервиса была использована ER-модель.

ER-модель (от англ. Entity-Relationship model, модель «сущность — связь») – модель данных, позволяющая описывать концептуальные схемы предметной области. ER-модель используется при высокоуровневом (концептуальном) проектировании баз данных. Она позволяет выделить ключевые сущности и отобразить связи – семантические отношения, которые могут устанавливаться между этими сущностями и чаще всего выражаются с помощью глаголов.

Модель была предложена в 1976 году Питером Ченом, им же предложена и самая популярная графическая нотация для модели (нотация П. Чена [1]). Однако в данном

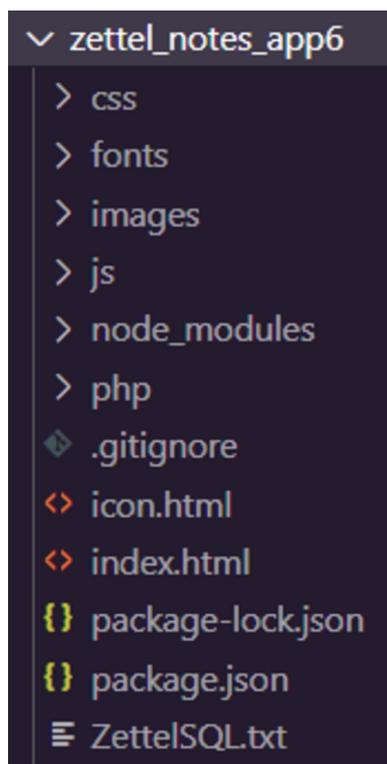


Рисунок 1. Файловая структура проекта

проекте будет использована другая нотация – нотация «Crow’s foot» [2], предложенная Гордоном Эверестом и берущая начало в его статье «Basic Data Structure Models Explained With A Common Example» [3].

Ниже представлена принципиальная и детализированная схемы ER-модели базы данных веб-приложения, выполненная с помощью веб-сервиса <https://app.diagrams.net/>.

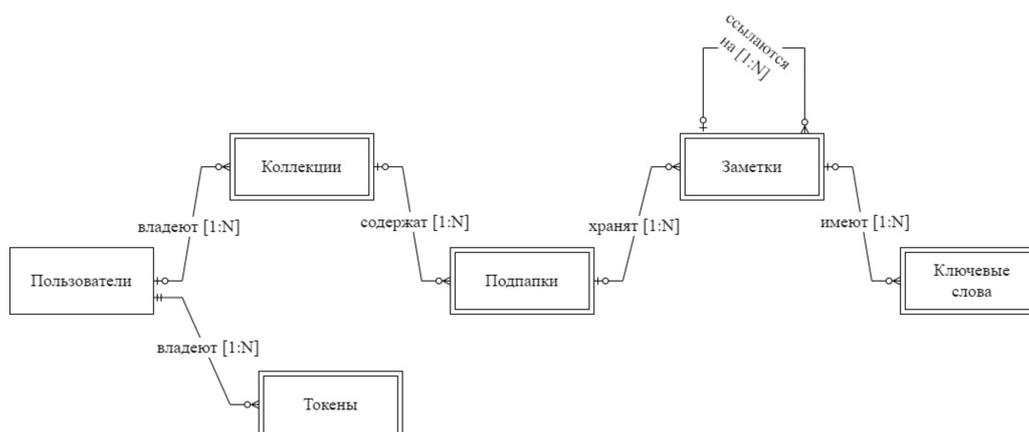


Рисунок 2. Принципиальная диаграмма ER-модели

Для начала работы в приложении пользователю необходимо создать учётную запись, заполнив поля формы регистрации.

После того, как учётная запись создана, а адрес почты подтверждён, клиент может войти в аккаунт, введя свои данные в поля формы авторизации.

Войдя в аккаунт, пользователь получает доступ к интерфейсу работы с заметками.

В приложении подключён модуль текстового редактора Quill-js, написанный на языке JavaScript [4]. В рамках нашего проекта он позволяет осуществлять несложное форматирование текста, добавлять изображения в качестве содержимого заметок, а также вставлять

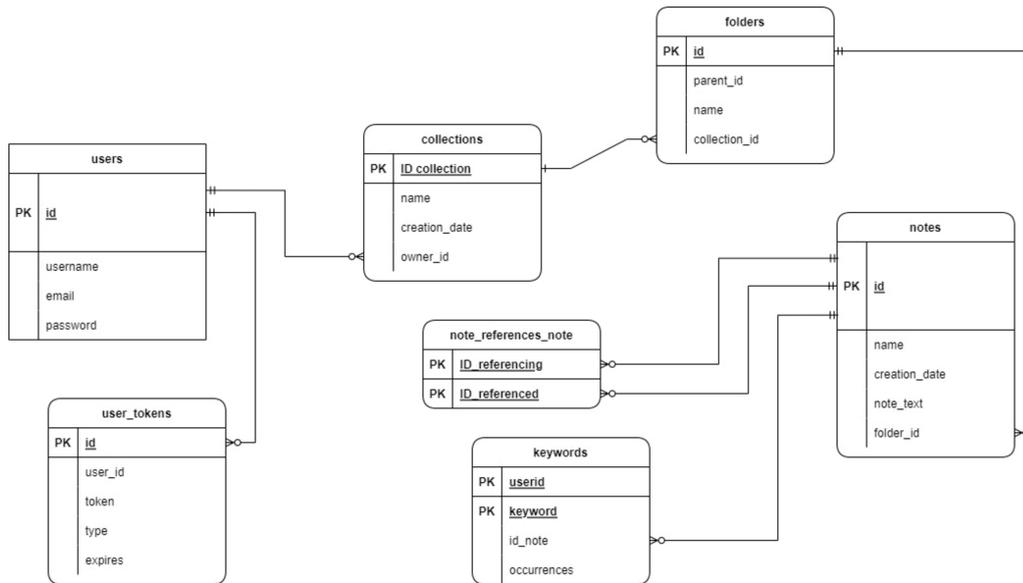


Рисунок 3. Детализированная ER-модель

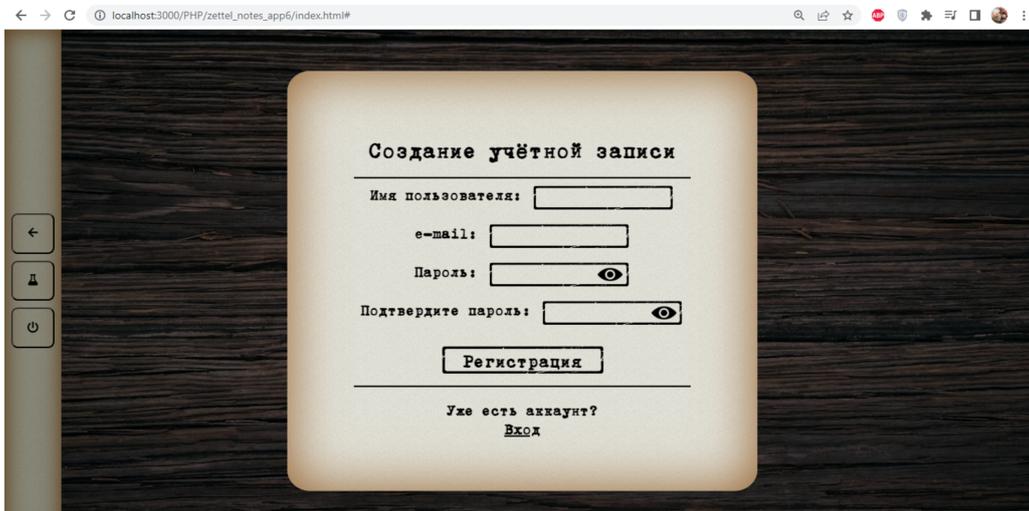


Рисунок 4. Форма регистрации новой учётной записи

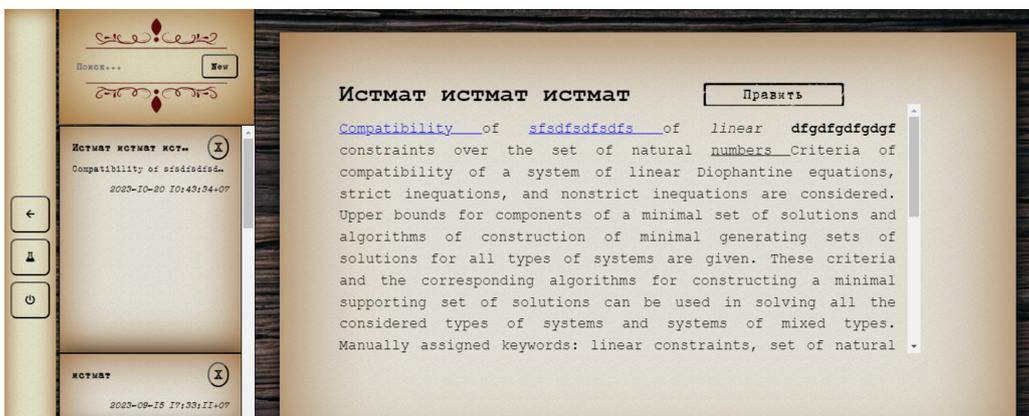


Рисунок 5. Общий вид окна приложения

в тексты записей ссылки как на внешние ресурсы, так и на другие заметки пользователя. Конечно, реальный потенциал Quill-js представляет из себя нечто большее, чем те задачи, для которых он используется в приложении на данный момент – к примеру, в перспективе

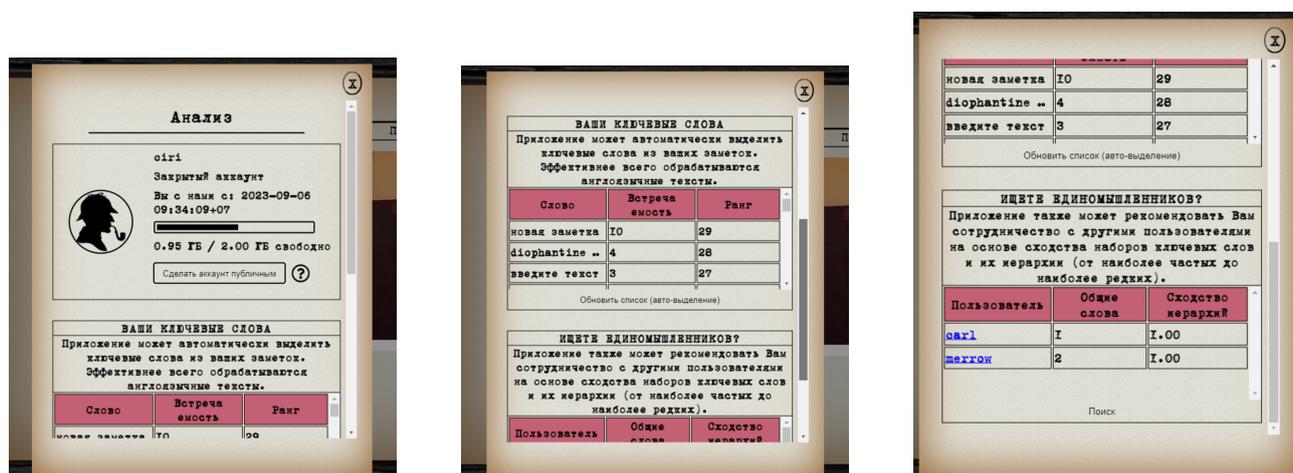
возможно внедрение возможности вставлять в текст заметки формулы, для чего также потребуется модуль по типу KaTeX.

У самого левого края окна расположена панель с тремя кнопками. Работа самой верхней из них на данный момент не реализована, однако в будущем предполагается, что она будет отвечать за сворачивание и разворачивание панели со всеми заметками пользователя и поисковой формой (расположена правее). Это сделает приложение более адаптивным и удобным для использования на экранах размером поменьше. Следующая кнопка (со значком колбы) – это кнопка вызова интерфейса работы с ключевыми словами, к чему мы вернёмся ниже. Самая нижняя кнопка отвечает за выход из аккаунта (при её нажатии страница перезагружается, а на экране появляется форма авторизации).

В левом верхнем углу находятся поисковая строка и кнопка создания новой заметки («New»), ниже – элементы предварительного просмотра всех имеющихся в коллекции заметок, информация о которых приходит из базы данных приложения. В правой части экрана – главное поле работы с отдельной заметкой, открывающееся по нажатию на соответствующий элемент предварительного просмотра. Любые изменения заголовка или содержания заметки в этом поле отражаются на данных о записке, хранимых в БД. Обновление данных в базе данных и в боковой панели происходит без перезагрузки страницы, как только пользователь «переключается» с редактируемого элемента на любой другой элемент страницы.

Кроме того, каждый элемент предварительного просмотра содержит в себе кнопку «X» – по её нажатию записка может быть удалена из базы данных, а её элемент предварительного просмотра исчезает с экрана. Если заметка была открыта для чтения/редактирования, она закрывается перед удалением.

Информация о пользователях и их заметках хранится в БД. В приложении реализован механизм ограничения количества памяти, которая может быть выделена под записи одного пользователя (для рабочей версии приложения этот лимит равен 2 Гб на одного клиента).



а) б) в)
 Рисунок 6. Рабочая версия интерфейса работы с ключевыми словами: а) секция с общей информацией об учётной записи пользователя; б) секция ключевых слов; в) секция возможных сотрудников

По нажатию на кнопку со значком колбы открывается окно работы с ключевыми словами и поиска сотрудников. На рисунке 6 а) показана секция с общей информацией об аккаунте пользователя. Важно упомянуть, что процедура поиска товарищей по интересам доступна только пользователям, чьи данные о ключевых словах и адресе электронной

почты являются открытыми. Сделать эту информацию открытой можно, нажав на кнопку «Сделать аккаунт публичным», после чего надпись изменится на «Сделать аккаунт закрытым» – и доступ к этим данным можно будет снова ограничить в любой момент.

Для запуска процедуры выделения ключевых слов необходимо нажать на кнопку «Обновить список» в нижней части секции «Ваши ключевые слова» (рисунок 6 б). Информация о выделенных ключевых словах будет сохранена в БД в таблице “keywords”. Если на момент входа в учётную запись пользователь уже запускал эту процедуру в прошлом, ключевые слова будут отображены в таблице.

После выделения ключевых слов пользователь может запустить поиск возможных сотрудников. Для этого ему нужно нажать на кнопку «Поиск» в нижней части секции «Ищете единомышленников?». Стоит заметить, что всякое имя пользователя в этой таблице является ссылкой на доступный для копирования адрес электронной почты соответствующего клиента. С рабочим вариантом приложения можно ознакомиться по ссылке <https://github.com/LYuree/SherlockZettelNotes>.

3. Математическая модель приложения

Теперь, когда мы рассмотрели интерфейс выделения ключевых слов и поиска возможных товарищей по работе, стоит подробнее рассмотреть математические алгоритмы, использованные для решения этих задач.

По какому принципу приложение выделяет ключевые слова? Для этой задачи было решено применить алгоритм RAKE. Алгоритм RAKE (Rapid Automatic Keywords Extraction) – один из алгоритмов выделения ключевых слов из фрагментов текста, описанный в книге “Text Mining: Applications and Theory” (Michael W. Berry, Jacob Kogan) [5]. Основной целью при разработке алгоритма стало нахождение метода извлечения ключевых слов, который был бы эффективен и легко применим к разным типам документов, не обязательно подчиняющихся строгим грамматическим конвенциям [6].

Работа алгоритма основывается на том, что ключевые слова зачастую оказываются в окружении пунктуации и т.н. стоп-слов (предлогов, союзов, частиц и т.д.), не несущих самостоятельной смысловой нагрузки и использующихся только в сочетании с другими словами [6, с. 4]. Используя стоп-слова как разделители, алгоритм разбивает текст на слова и фразы-кандидаты [6, с. 5]. Далее кандидаты ранжируются с использованием следующих метрик [6, с. 6]: частота встречаемости слова в тексте заметки ($freq(w)$), word degree ($deg(w)$) и отношение degree к частоте:

$$score(w) = \frac{deg(w)}{freq(w)}.$$

Для большей наглядности возьмём отрывок научного текста с ключевыми словами, присвоенными вручную: *“Compatibility of systems of linear constraints over the set of natural numbers Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types of systems and systems of mixed types. Manually assigned keywords: linear constraints, set of natural numbers, linear Diophantine equations, strict inequations, nonstrict inequations, upper bounds, minimal generating sets”*.

Ключевые слова, выделенные вручную: *linear constraints, set of natural numbers, linear Diophantine equations, strict inequations, nonstrict inequations, upper bounds, minimal generating sets.*

Кандидаты в ключевые слова, согласно алгоритму RAKE: *Compatibility, systems, linear constraints, set, natural numbers, Criteria, compatibility, system, linear Diophantine equations, strict inequations, nonstrict inequations, Upper bounds, components, minimal set, solutions, algorithms, minimal generating sets, solutions, systems, criteria, corresponding algorithms, constructing, minimal supporting set, solving, systems, systems.*

Следующим шагом алгоритма является построение матрицы со-встречаемости (в оригинале статьи вместо термина «матрица» употреблён термин «граф» – «co-occurrence graph») слов – назовём эту матрицу C . Элемент матрицы C_{ij} – это частота встречаемости i -го слова в одной фразе с j -м словом. Понятно, что элемент C_{ij} (на главной диагонали) будет равен частоте встречаемости слова в принципе (со-встречаемости слова в одной связке с самим собой).

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
algorithms	2						1																		
bounds		1																							1
compatibility			2																						
components				1																					
constraints					1								1												
constructing						1																			
corresponding	1						1																		
criteria								2																	
diophantine									1	1			1												
equations									1	1			1												
generating											1		1						1						
inequations												2				1					1				
linear				1					1	1			2												
minimal										1				3				2	1			1			
natural															1		1								
nonstrict												1				1									
numbers															1		1								
set														2				3				1			
sets										1			1						1						
solving																				1					
strict											1										1				
supporting																						1			
system																							1		
systems																								4	
upper		1																							1

Рисунок 7. Матрица со-встречаемости для кандидатов

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

Рисунок 8. Очки выделенных слов

Из рисунков 7 и 8 видно, что $freq(w)$ – это частота встречаемости слова в принципе, а $deg(w)$ определяется как сумма $freq(w)$ и частот встречаемости слова с другими словами, так что $deg(w)$ всегда больше или равно $freq(w)$. Отсюда видно, что кандидаты, чаще встречающиеся вместе с другими словами, получают больший балл в отношении $deg(w) / freq(w)$.

После того, как баллы кандидатов рассчитаны, верхние T кандидатов отбираются в качестве ключевых слов (например, треть). В приведённом выше отрывке научного текста

28 ключевых слов, $T = 9$. На рисунке 9 сравниваются набор ключевых слов, отобранных с помощью RAKE, и набор слов, отобранных вручную.

Extracted by RAKE	Manually assigned
minimal generating sets	minimal generating sets
linear diophantine equations	linear Diophantine equations
minimal supporting set	
minimal set	
linear constraints	linear constraints
natural numbers	
strict inequations	strict inequations
nonstrict inequations	nonstrict inequations
upper bounds	upper bounds
	set of natural numbers

Рисунок 9. Сравнение двух наборов ключевых слов

График на рисунке 10 демонстрирует эффективность работы алгоритма в зависимости от размера анализируемого текста в сравнении с другим алгоритмом выделения ключевых слов, TextRank.

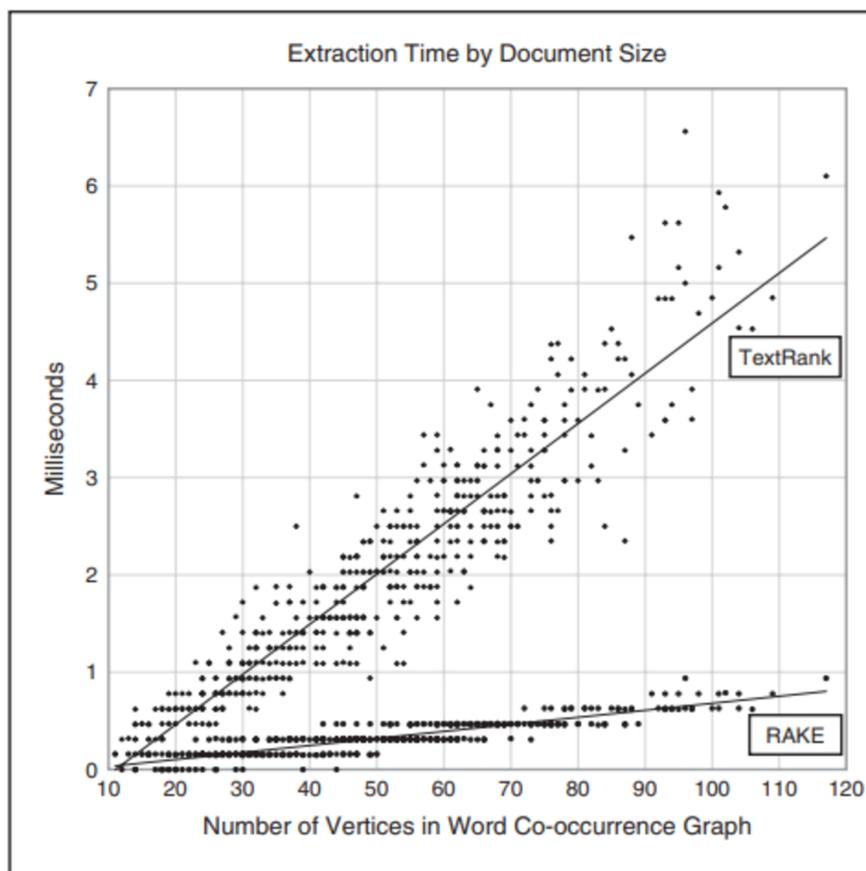


Рисунок 10. Время работы алгоритмов RAKE и TextRank в миллисекундах (ms) в зависимости от кол-ва вершин в графе со-встречаемости слов

Что касается реализации алгоритма, в программном модуле rake-php-plus (см. ссылку на удалённый репозиторий github [7]) разработчик добавил файлы со списками стоп-слов для 16 различных языков, в т.ч. русского. По нажатию кнопки «обновить список» данные каждой заметки отправляются на сервер с помощью объекта XMLHttpRequest, после чего

серверный скрипт на языке PHP применяет к тексту заметки алгоритм RAKE и возвращает массив слов обратно на клиентское устройство, где с помощью языка JS ключевые слова выводятся в таблицу. Слова сохраняются в базе данных, так что при перезагрузке страницы браузер, автоматически отправляя запрос на сервер, получает от него массив ключевых слов и отображает их в таблице.

Выделение ключевых слов, таким образом, происходит на стороне сервера с помощью языка PHP.

Что же касается поиска потенциальных сотрудников, в этой задаче применяется следующий порядок действий. Набор ключевых слов пользователя, запустившего процедуру поиска, поочередно сравнивается с наборами ключевых слов других пользователей, открывших доступ к своим данным. Таким образом находится массив общих ключевых слов (пересечение множеств). Для обоих пользователей – для клиента, запустившего поиск, и для сравниваемого с ним пользователя – эти слова ранжируются, при этом учитываются ранги, которые были назначены этим словам при изначальном их выделении из текстов. В результате мы получаем две ранговые системы ключевых слов – с одними и теми же, общими для двух людей ключевыми словами, но с разными (в общем случае) рангами. Показатель корреляции между этими ранговыми системами предлагается использовать в качестве меры сходства интересов пользователей. Внимательный читатель, однако, сразу же заметит в этом подходе явный недостаток, который достаточно быстро обнаружил себя при разработке приложения, но об этом будет сказано ниже. Для вычисления коэффициента корреляции применяется формула коэффициента корреляции Спирмена [8]:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}.$$

4. Тестирование

Пример работы алгоритма RAKE.

Создадим для одного из пользователей заметку, текст которой будет содержать материал научной статьи “Prevention and Control of Air Pollution in China: A Research Agenda for Science and Technology Studies” [9], и запустим процедуру выделения ключевых слов.

Ключевые слова, указанные авторами статьи: *air pollution, China, 12th FYP, Science and Technology Studies, STS.*

Ключевые слова, выделенные с помощью программной реализации алгоритма RAKE: *introduction in recent years, china’s air pollution has attracted the world’s attention for the images of its main cities shrouded in haze and heavy smog, in fact, a recent official report1 shows that in 2014, only 8 of 74 cities monitored by china’s ministry of environmental protection met air quality standards, the origins of china’s modern air pollution problem can be linked to the “reform and open policy” of 1978, which introduced a new era of rapid and ambitious economic development, during more than three decades, continued economic growth has been accomplished at the expense of an excessive use of energy that has placed an enormous burden on the ecological system, fang et al, parallel to the evolution of its economic policy, china’s environmental governance, and more specifically, its air pollution management, were adopted in the decade of the 1970s and have since been continuously updated, large reforms on the modernization of environmental protection institutions, the standards to control pollutants, and the work on legal framework, among others, have expressed the will of the central government to advance in the prevention and control of air pollution, pcap, more recently, the 12th five-year plan, fyp, of 2011-2015, which sets out both economic and political guidelines, represents a change of vision regarding environmental pollution as a problem related to the entire system, wong, prevention and control of air pollution in china: a research agenda for science and technology studies.*

Из результатов видно, что программная реализация алгоритма часто даёт на выходе фразы, слишком длинные по сравнению с теми, что обычно указываются авторами в качестве ключевых слов. К тому же, из текста было выделено много слов и фраз, которые вряд ли можно считать ключевыми, исходя из их важности для отрывка. Отсюда, как нам видится, можно сделать несколько выводов:

- вероятно, полезным усовершенствованием приложения будет предоставление пользователю возможности назначать ключевые слова своим заметкам вручную; система автоматического выделения ключевых слов при этом может использоваться в качестве вспомогательного средства, дающего рекомендации;
- ещё одним возможным улучшением могла бы стать замена системы алгоритма RAKE на одну из существующих сегодня текстовых нейросетей, использующих технологию Natural Language Processing (NLP), в качестве средства выделения ключевых слов.

Перейдём к примеру работы алгоритма вычисления коэффициента корреляции. Рассмотрим три ранжированных набора ключевых слов.

Набор 1:

Слово	Ранг
compatibility	21
systems	20
linear constraints	19
set	18
natural numbers	17
criteria	16
system	15
linear diophantine equations	14
strict inequations	13
nonstrict inequations	12
upper bounds	11
components	10

Слово	Ранг
minimal set	9
solutions	8
algorithms	7
minimal generating sets	6
criteria	5
corresponding algorithms	4
constructing	3
minimal supporting set	2
solving	1

Набор 2:

Слово	Ранг
blood vessels	6
heart	5
heart diseases	4
diseases	3
circulatory	2
lymphatic	1

Набор 3:

Слово	Ранг
systems	27
linear constraints	26
set	25
natural numbers	24
criteria	23
compatibility	22
system	21
linear diophantine equations	20
strict inequations	19
nonstrict inequations	18
upper bounds	17
components	16
minimal set	15
solutions	14

Слово	Ранг
algorithms	13
minimal generating sets	12
criteria	11
corresponding algorithms	10
constructing	9
minimal supporting set	8
solving	7
blood vessels	6
heart	5
heart diseases	4
diseases	3
circulatory	2
lymphatic	1

Набор 4:

Слово	Ранг
systems	7
blood vessels	6
heart	5
heart diseases	4
diseases	3
circulatory	2
lymphatic	1

При подсчёте коэффициента корреляции для пары из двух наборов ключевых слов нужно учесть тот факт, что число элементов в изначальных наборах может быть неодинаковым, а значит ранги для общих слов могут находиться в совершенно разных диапазонах. Поэтому после того, как найдено пересечение двух множеств, нужно составить два набора из общих ключевых слов, где словам будут присвоены новые, в общем случае разные ранги согласно тому порядку, в котором эти словам находились в исходных массивах.

Матрица коэффициентов корреляции для пересечений множеств этих наборов будет иметь следующий вид:

	a1	a2	a3	a4
a1	1	0	0.89	1
a2	0	1	1	1
a3	0.89	1	1	1
a4	1	1	1	1

У 1-го и второго наборов нет общих слов, поэтому коэффициент корреляции Спирмена для них равен 0. У 1-го и 3-го наборов есть общая часть:

Слово	Ранг по набору a1	Ранг по набору a3
compatibility	21	16
systems	20	21
linear constraints	19	20
set	18	19
natural numbers	17	18
criteria	16	17
system	15	15
linear diophantine equations	14	14
strict inequations	13	13
nonstrict inequations	12	12
upper bounds	11	11
components	10	10
minimal set	9	9
solutions	8	8
algorithms	7	7
minimal generating sets	6	6
criteria	16	17
corresponding algorithms	4	4
constructing	3	3
minimal supporting set	2	2
solving	1	1

Коэффициент корреляции равен 0.89, что является высоким значением.

5. Сложности, возникшие в процессе работы над проектом

Говоря о сложностях, возникших при реализации проекта, нельзя не отметить задачу выбора критерия, на основании которого можно было бы давать пользователям небезосновательные рекомендации о сотрудничестве друг с другом. Проблема метода определения степени сходства интересов пользователей на основе корреляции ранговых систем их общих ключевых слов, очевидно, состоит в том, что, даже если уровень корреляции между ранжированными наборами общих ключевых слов окажется высоким, сама доля совпавших ключевых слов для двух пользователей при этом может быть ничтожно мала. В этом случае показатель корреляции способен привести нас к ложным выводам о сильном сходстве интересов клиентов. Одна из перспектив развития приложения – выбор лучшего показателя сходства наборов ключевых слов из заметок пользователей. В качестве альтернативы показателю корреляции в данный момент рассматривается критерий сходства множеств, а также метод оценки с помощью апостериорного максимума (MAP).

Список литературы

1. Чен П.П.-Ш. Модель “сущность-связь” — шаг к единому представлению о данных / пер. с англ. М.Р. Коголовский // Системы управления базами данных. — 1995. — № 3. — С. 137–158.
2. Crow’s Foot Notation. — URL: <http://www2.cs.uregina.ca/~bernatja/crowsfoot.html>. Дата обращения: 20.12.2022.
3. Everest G. Basic Data Structure Models Explained With A Common Example // Proceedings Fifth Texas Conference on Computing Systems, Austin, TX, 1976 October 18–19. — Long Beach. — CA : IEEE Computer Society Publications Office, 1976. — P. 39–42.

4. Удалённый репозиторий программного модуля quill-js. — URL: <https://github.com/quilljs>. Дата обращения: 06.11.2023.
5. Michael W. Berry, Kogan J. Text Mining: Applications and Theory. — URL: <https://onlinelibrary.wiley.com/doi/book/10.1002/9780470689646>. Дата обращения: 14.06.2023.
6. RAKE (Rapid Keyword Extraction Algorithm). — URL: https://www.researchgate.net/publication/227988510_Automatic_Keyword_Extraction_from_Individual_Documents/link/59edf51fa6fdccbbefd5434a/download. Дата обращения: 13.06.2023.
7. Программный модуль RAKE-PHP. — URL: <https://github.com/Donatello-za/rake-php-plus>. Дата обращения: 13.06.2023.
8. Corder G.W., Foreman, D.I. Nonparametric Statistics: A Step-by-Step Approach. — NY : Wiley, 2014.
9. Rodolfo Andres Hernandez. Prevention and Control of Air Pollution in China: A Research Agenda for Science and Technology Studies // S.A.P.I.E.N.S. — URL: <http://journals.openedition.org/sapiens/1734>.