При необходимости, отдельные базы данных могут быть с лёгкость объединены в одну.

Важно отметить, что линейный список AST-строк для каждой лабораторной работы по мере наполнения базы данных возрастает, алгоритмическая сложность поиска в нем составляет O(n). Но за счёт древовидной структуры представленной выше базы данных снижается скорость нахождения нужного списка с AST-строками, что оптимизирует работу данной системы.

Разработанная система обнаружения плагиата учитывает различные способы видоизменения программного кода, написанного на языке Python. Система протестирована на работоспособность, наилучшую точность обнаружения плагиата показал алгоритм сравнения программных кодов, предварительно преобразованных в AST-строки.

Список литературы:

1. Пунчик, В.Н. Поликонтекстный анализ феномена «плагиат» в информационном обществе / В.Н. Пунчик, З.В. Пунчик // Социология – 2016. – № 1. – С. 83-91.

УДК 004.43+81.33+026.06

БИБЛИОТЕКА ЭЛЕКТРОННЫХ ДОКУМЕНТОВ С ВОЗМОЖНОСТЬЮ ИСПОЛЬЗОВАНИЯ В РЕЖИМЕ БАЗЫ ЗНАНИЙ С САМООБСЛУЖИВАНИЕМ

Я. А. Примак
Гродненский государственный университет имени Янки Купалы, Гродно
Научный руководитель: А. М. Кадан, кандидат технических наук, доцент

Базы знаний с самообслуживанием в настоящее время набирают популярность. Под этим термином подразумевают централизованный и структурированный сборник информации определенного направления - о продукте, услуге, отделе или теме. Иногда такую базу называют базой знаний по обслуживанию клиентов или базой знаний службы поддержки клиентов. Часто она имеет возможности поиска и содержит практические советы и инструкции, которые помогают клиентам — внутренним или внешним — решать проблемы, не обращаясь в службу поддержки.

Следуя указанной тенденции, в рамках междисциплинарного проекта в Гродненском государственном университете имени Янки Купалы было решено создать библиотеку электронных документов с возможностью использования в режиме базы знаний с самообслуживанием.

В работе использованы принципы построения баз знаний с самообслуживанием, а также реализован прототип поискового алгоритма. Его особенность в том,

что он обеспечивает возможности более точного поиска, в том числе и по содержанию книги, преобразованному согласно технологиям NLP (обработки естественных языков).

В состав электронной библиотеки входят различные разделы, включая каталоги, списки литературы, метаданные, краткие описания и полные тексты документов.

Так как библиотека является ведомственной (университетской), то доступ к ее сайту и материалам предоставляется только авторизованным пользователям, зарегистрированным пользователям, зарегистрированным в едином информационном пространстве университета. Также, из-за требований соблюдения авторского права, доступ к материалам сайта разрешен только из локальной сети университета.

Согласно концепции решения на сайте электронной библиотеки присутствуют такие роли пользователей, как Студент, Преподаватель и Модератор.

Основные функциональные возможности группы пользователей **Студент**:

- Искать книгу,
- Просматривать информацию о книге,
- Скачать (читать) книгу,
- Ставить лайк и/или оставлять комментарий,
- Формировать список Избранное.

Основные функциональные возможности группы пользователей **Преподаватель** такие же как у группы **Студент**. За исключением возможности «Формировать подборки книг для подготовки студентов». Также имеет возможность просматривать статистику о пользователях и о рейтинге книг (как часто просматривали, скачивали, количество лайков, комментариев).

Модератор, кроме указанного выше, может совершать все CRUD-операции с книгами.

Для разработки электронной библиотеки была выбрана платформа CMS Joomla, так как она имеет все необходимые функции для управления контентом, пользовательскими аккаунтами и доступом, а также широкие возможности для настройки и расширения функционала [1]. Соответственно, для хранения данных проекта используется СУБД MySQL.

В традиционной библиотеке поиск необходимого издания ведется по конечному набору реквизитов «библиотечной карточки» (автор, название, издательство, год выпуска и т.п.). Информация из «библиотечной карточки», содержащей краткое, в 1-2 предложения, описание содержания книги используется пользователем только в ознакомительных целях, не участвуя в поисковом запросе и позволяя пользователю более точно выбрать нужную книгу из предложенной поисковой выдачи.

Из вышеописанного следует, что атрибутов книги, традиционно используемых для поиска, категорически недостаточно для решения задачи поиска в базе данных с самообслуживанием. Важно выделить и структурировать дополнительные данные о книгах, авторах, жанрах и оглавлении таким образом, чтобы они были легко доступны и индексируемы для поиска.

Ключевой особенностью проекта является «умный» поиск, реализация которого включает применение систем интеллектуального анализа текстов.

Предварительная подготовка информации предполагает Извлечение оглавления электронного издания, которое включает сканирование (фотографирование) оглавления, распознавание текста оглавления, очистку и форматирование полученного текста. Контекстная единица в этом случае – содержание главы.

Также в ходе предварительной подготовки информации оглавления выполняется токенизация по предложениям, токенизация по словам, удаление стоп-слов, лемматизация и стемминг текста. Кратко их суть в следующем. **Удаление стоп-слов** - иногда одних слов в тексте больше, чем других, к тому же они встречаются почти в каждом предложении и не несут большой информативной нагрузки. Такие слова являются шумом для последующего глубокого обучения и называются стоп-словами. **Стемминг –** уменьшает разнообразие морфологической структуры информации. Так слова «начальный» и «начальное» имеют тот же смысл, но разную форму, например, «начальное значение» и «начальное приближение». Поэтому для машинного обучения и анализа лучше привести их к одной форме для уменьшения размерности. В частности, стемминг опускает окончания слова. **Лемматизация -** над словом проводится морфологический анализ с целью выявить его начальную форму. Например, «хочу», «хотят», «хотели» сводятся к начальной форме «хотеть» [3].

Как и задачи нормализации текста, задачи построения n-грамм также решаются с использованием методов на основе технологий NLP (Natural Language Processing, обработка естественных языков) [2].

Для дальнейшей обработки производится сборка содержимого контекстных единиц, предполагающая построение связных структур, объединяющих название книги, название главы, название параграфа, название раздела параграфа. Поиск в режиме базы знаний с самообслуживанием будет производиться алгоритмами нахождения в пространстве структур оглавления минимального расстояния между запросом пользователя и данными оглавлений электронных книг.

В проекте, в дополнении к NLP, используется технология векторного представления слов — **word2vec**. Word2vec позволяет получить и использовать числовые представления слов, рассматривая слова, окружающие данное слово. Он изучает значение слова, просматривая его контекст и представляя его численно. Основная идея Word2Vec заключается в том, чтобы преобразовать слова в векторы в n-мерном пространстве таким образом, чтобы близкие по смыслу слова имели близкие векторные представления [4].

Для организации поиска, наряду и традиционным методом сравнения запроса и данных оглавлений, используются несколько методов, в том числе методы, основанные на сравнении строк, на вычислении расстояния Левенштейна, а также методы, основанные на технологии векторного представления слов.

В заключение необходимо отметить, что электронная библиотека становится важным элементом современной образовательной среды факультета математики и информатики Гродненского государственного университета им. Янки Купалы, обеспечивая удобство и доступность современной ИТ-литературы для студентов и преподавателей.

Список литературы

1. Обзор функционала Joomla! [Электронный ресурс] / Сайт проекта Joomla!. – URL: https://joomla.ru/docs/articles/cms-joomla/1821-joomla-opportunities (дата обращения: 08.11.2023).

2. Малюшкин, Р. NLP для людей. Часть 1 [Электронный ресурс] / Р. Малюшкин // Medium – Where good ideas find you. – URL: https://medium.com/stseusp/nlp-for-people-1-c9b54ffce13f (дата обращения: 09.11.2023).

3. Котюбеев, Р. Предобработка текста в NLP [Электронный ресурс] / Р. Котюбеев // PYTHON SCHOOL. – URL: https://python-school.ru/blog/nlp-text-preprog716754540 (дата обращения: 09.11.2023).

4. Mikolov, Tomas; et al. Efficient Estimation of Word Representations in Vector Space [Электронный ресурс] / arXiv.org e-Print archive. – URL: https://arxiv.org/pdf/1301.3781 (дата обращения: 08.11.2023).

УДК 004.8

ПОСТРОЕНИЕ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ С ПОМОЩЬЮ БИБЛИОТЕКИ PYTORCH

И. В. Савицкий
“Гродненский государственный университет им. Янки Купалы”, г.Гродно
Научный руководитель: И. Б. Просвирнина, кандидат физико-математических наук, доцент

Сверточные нейронные сети (CNN) являются мощным инструментом для анализа и распознавания изображений. Они могут быть применены во множестве сфер человеческой жизни, где требуется классификация объектов или процессов по определенным критериям. Особый интерес представляет их применение в медицине, где большой объем данных и сложные взаимодействия между переменными способствуют эффективной моделированию с помощью глубокого обучения. Использование CNN в анализе медицинских изображений и сигналов позволяет достичь высокой точности диагностики и улучшить качество медицинской практики.

Для поиска наборов данных будем использовать платформу Kaggle. Выберем датасет Chest X-Ray Images (Pneumonia), организованный в 3 папки (train, test, val) и содержащий подпапки для каждой категории изображений (пневмония/норма). Имеется 5863 рентгеновских изображения (JPEG) и 2 категории (пневмония/норма).