Russian Technological Journal.
2024;12(5):7–16

ISSN 2782-3210 (Print)
ISSN 2500-316X (Online)

Информационные системы. Информатика. Проблемы информационной безопасности

Information systems. Computer sciences. Issues of information security

УДК 025.4.03
https://doi.org/10.32362/2500-316X-2024-12-5-7-16
EDN CBEERK

НАУЧНАЯ СТАТЬЯ

Автоматизация поиска юридической информации на арабском языке: подход к поиску документов

Камел С. Жафар [1, @],
Али А. Мохаммад [2],
Али Х. Исса [3],
А.В. Панов [1, @]

[1] *МИРЭА – Российский технологический университет, Москва, 119454 Россия*
[2] *Национальный исследовательский университет «Высшая школа экономики», Москва, 109028 Россия*
[3] *РОСБИОТЕХ – Российский биотехнологический университет, Москва, 125080 Россия*
[@] *Авторы для переписки, e-mail: zhafar.k@edu.mirea.ru, panov_a@mirea.ru*

Резюме

Цели. Поиск юридической информации, например, информации, связанной с различными юридическими вопросами, такими как наказание за преступления, является сложной задачей. Предлагаемый авторами подход может быть эффективным и действенным способом автоматизации поиска юридической информации без необходимости использования большого количества размеченных данных или значительных вычислительных ресурсов. Целью статьи является анализ возможности использования подхода к поиску документов в контексте юридических текстов на арабском языке, с применением методов обработки естественного языка и неконтролируемой кластеризации.

Методы. Использован подход Top2Vec – алгоритм моделирования темы, который создает вложения документов на основе семантического контекста, чтобы группировать юридические тексты на арабском языке в соответствующие темы. Использован алгоритм кластеризации на основе плотности для определения подтем внутри каждого кластера. Решаются проблемы работы с арабским юридическим текстом, такие как морфологическая сложность, двусмысленность и отсутствие стандартизированной терминологии. Предложен конвейер предварительной обработки, включающий в себя токенизацию, нормализацию, выделение корней и удаление стоп-слов.

Результаты. Результаты оценки подхода с использованием набора данных юридических текстов на арабском языке, основанного на ключевых словах, показали его эффективность и превосходство с точки зрения точности и запоминаемости. Предлагаемый подход обеспечивает точность поиска – 87% и полноту поиска – 80%. Применение этого подхода может значительно улучшить поиск юридических документов, сделав его более быстрым и точным.

Выводы. Предложенный подход может быть ценным инструментом для юристов и исследователей, которым необходимо ориентироваться в обширном и сложном ландшафте арабской юридической информации, повышая эффективность и точность ее поиска.

Ключевые слова: поиск документов, обработка естественного языка, Top2Vec, алгоритм кластеризации на основе плотности, арабские юридические документы, вложения слов, косинусное сходство

• **Поступила:** 05.05.2023 • **Доработана:** 04.04.2024 • **Принята к опубликованию:** 11.07.2024

Для цитирования: Жафар К.С., Мохаммад А.А., Исса А.Х., Панов А.В. Автоматизация поиска юридической информации на арабском языке: подход к поиску документов. *Russ. Technol. J*. 2024;12(5):7–16. https://doi.org/10.32362/2500-316X-2024-12-5-7-16

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

RESEARCH ARTICLE

Automating the search for legal information in Arabic: A novel approach to document retrieval

Kamel S. Jafar [1, @], Ali A. Mohammad [2], Ali H. Issa [3], Alexander V. Panov [1, @]

[1] *MIREA – Russian Technological University, Moscow, 119454 Russia*
[2] *Higher School of Economics, Moscow, 109028 Russia*
[3] *Russian Biotechnological University, Moscow, 125080 Russia*
[@] *Corresponding authors, e-mail: zhafar.k@edu.mirea.ru, panov_a@mirea.ru*

Abstract

Objectives. The retrieval of legal information, including information related to issues such as punishment for crimes and felonies, represents a challenging task. The approach proposed in the article represents an efficient way to automate the retrieval of legal information without requiring a large amount of labeled data or consuming significant computational resources. The work set out to analyze the feasibility of a document retrieval approach in the context of Arabic legal texts using natural language processing and unsupervised clustering techniques.

Methods. The Topic-to-Vector (Top2Vec) topic modeling algorithm for generating document embeddings based on semantic context is used to cluster Arabic legal texts into relevant topics. We also used the HDBSCAN density-based clustering algorithm to identify subtopics within each cluster. Challenges of working with Arabic legal text, such as morphological complexity, ambiguity, and a lack of standardized terminology, are addressed by means of a proposed preprocessing pipeline that includes tokenization, normalization, stemming, and stop-word removal.

Results. The results of the evaluation of the approach using a dataset of legal texts in Arabic based on keywords demonstrated its superior effectiveness in terms of accuracy and memorability. The proposed approach provides 87% accuracy and 80% completeness. This circumstance can significantly improve the search for legal documents, making the process faster and more accurate.

Conclusions. Our findings suggest that this approach can be a valuable tool for legal professionals and researchers to navigate the complex landscape of Arabic legal information to improve efficiency and accuracy in legal information retrieval.

Keywords: search for documents, NLP, Top2Vec, HDBSCAN, Arabic legal documents, word embeddings, cosine similarity

• **Submitted:** 05.05.2023 • **Revised:** 04.04.2024 • **Accepted:** 11.07.2024

For citation: Jafar K.S., Mohammad A.A., Issa A.H., Panov A.V. Automating the search for legal information in Arabic: A novel approach to document retrieval. *Russ. Technol. J.* 2024;12(5):7–16. https://doi.org/10.32362/2500-316X-2024-12-5-7-16

Financial disclosure: The authors have no financial or property interest in any material or method mentioned.

The authors declare no conflicts of interest.

ВВЕДЕНИЕ

Поиск и анализ юридической информации связаны с определенными трудностями, обусловленными спецификой области права. В последние годы наблюдается растущий интерес к использованию методов обработки естественного языка для автоматизации процесса доступа к такой информации [1]. В частности, были предприняты значительные усилия по разработке систем вопросов и ответов, которые могут извлекать конкретные ответы из юридических документов [2, 3]. Однако построение качественной системы контроля качества требует большого количества размеченных данных и часто связано со значительными вычислительными затратами. В работе предлагается альтернативный подход к автоматизации поиска юридической информации, связанной с преступлениями, в т.ч. уголовными, в нормативных актах и юридических документах на арабском языке. Во многих практических случаях достаточно получить наиболее релевантные юридические документы, связанные с запросом пользователя, без необходимости извлечения конкретных ответов из документов. Подход состоит из нескольких шагов, включая сбор данных, их предварительную обработку, индексацию документов, обработку запросов и поиск документов. В нашем наборе данных используются стандартные арабские грамматические и юридические документы, касающиеся преступлений и уголовных правонарушений.

1. СВЯЗАННЫЕ РАБОТЫ

Поиск юридической информации – сфера деятельности с богатой историей и обширным объемом исследований. Рассмотрим ключевые работы и разработки в этой области с особым акцентом на подходах к поиску юридической информации и документов на арабском языке. Авторы [4] исследуют современные методы искусственного интеллекта (ИИ), используемые для систем поиска правовой информации. С появлением информационных и коммуникационных технологий практикующие юристы столкнулись с резким ростом цифровой информации, что делает крайне важными эффективные методы поиска. В статье рассматриваются различные подходы к ИИ, включая методы обработки естественного языка, машинного обучения и извлечения знаний, а также то, как они могут помочь в поиске правовой информации. Кроме того, описываются проблемы, с которыми сталкиваются практикующие юристы, особенно при поиске аналогичных дел, законодательных актов или параграфов, и обсуждаются открытые вопросы, касающиеся систем поиска правовой информации. В целом исследование подчеркивает важность ИИ в правовой сфере и необходимость продолжения исследований и разработок в системах поиска правовой информации. В [5] представлены комментарии к девяти значимым источникам, опубликованным за последнее десятилетие. Четыре статьи посвящены анализу юридических дел, введению контекстуальных соображений, прогнозированию результатов на основе описаний дел на естественном языке, сравнению различных способов представления дел и формализации прецедентных рассуждений. Одна статья представляет метод анализа аргументов, который впоследствии стал очень широко использоваться в ИИ и праве, а именно – схемы аргументации. Две статьи относятся к онтологиям для представления правовых концепций, а две используют преимущества растущей доступности юридических наборов данных в этом десятилетии для автоматизации обобщения документов и поиска аргументов.

Авторы [6] представляют обзор развития, текущего состояния и будущих направлений правового ИИ. Юридический ИИ применяет обработку естественного языка, чтобы помочь юристам в их работе с потенциалом повышения эффективности за счет автоматизации утомительных задач. Авторы иллюстрируют точки зрения юристов и исследователей методов обработки естественного языка с помощью экспериментов и анализа существующих работ. Они определяют моделирование знаний, юридическое обоснование и интерпретируемость как три основные проблемы юридических задач, для решения которых требуются дальнейшие исследования. В документе предлагается объединить методы на основе встраивания и на основе символов для решения этих проблем, создания крупномасштабных и высококачественных наборов данных и учета этических проблем, таких как гендерная предвзятость и расовая

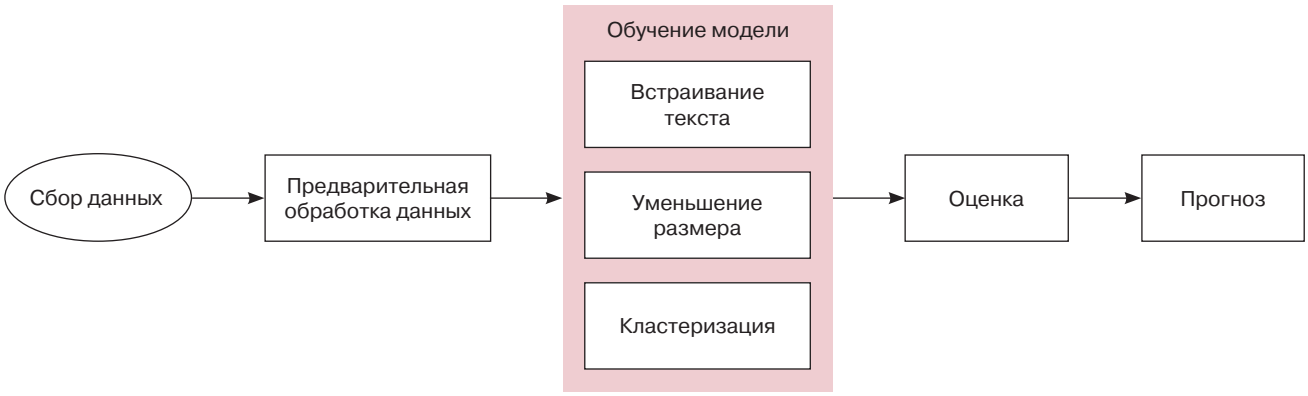

Рис. 1. Подход Top2Vec для поиска юридической информации

дискриминация. В конечном счете, юридический ИИ должен играть вспомогательную роль в правовой системе, а профессионалы должны принимать решения в первую очередь. Авторы [7] описывают разработку арабской системы извлечения информации из юридических документов, в которой используются гибридный подход машинного обучения и методы, основанные на правилах. Система предназначена для извлечения важной информации из документов и предоставления ее в структурированном виде для сложных запросов. Подход был протестирован на ограниченном классе арабских юридических документов и показал хорошие результаты. Авторы предлагают несколько возможных расширений системы, таких как рассмотрение различных типов дел в данной правовой системе, использование более продвинутых арабских инструментов методов обработки естественного языка, изучение использования глубокого обучения, извлечение большего количества отношений, улучшение представления результатов и расширение системы в другие области, такие как здравоохранение и финансы.

2. ПРЕДЛАГАЕМЫЙ ПОДХОД

Предлагаемая методология использования подхода Topic-to-Vector (Top2Vec) для поиска юридической информации содержит несколько шагов (рис. 1). Первым шагом является подготовка данных, которая включает сбор юридических документов, связанных с преступлениями и уголовными преступлениями, в арабских правилах и документах. Затем данные предварительно обрабатываются с использованием конвейера, который включает в себя токенизацию, нормализацию, выделение корней и удаление стоп-слов, чтобы подготовить текстовые данные для модели.

Следующим шагом является обучение модели, которое включает в себя создание вложений документов с использованием модели Top2Vec. Затем вложения сокращаются до пространства меньшей размерности с использованием методов уменьшения размерности, таких как UMAP[1], чтобы упростить кластеризацию документов. Кластеризация выполняется с использованием алгоритма кластеризации на основе плотности, например, HDBSCAN[2], который может эффективно идентифицировать кластеры документов на основе их сходства [8].

После завершения кластеризации модель можно оценить с помощью набора реальных пользовательских запросов. Для каждого запроса модель извлекает наиболее релевантные кластеры документов, и пользователь может просматривать документы, чтобы найти необходимую информацию. Производительность модели можно оценить с помощью таких показателей, как точность и полнота.

Наконец, модель можно использовать для прогнозирования, когда пользователь вводит запрос, а модель извлекает наиболее релевантные кластеры документов. Затем пользователь может просматривать документы, чтобы найти необходимую информацию. Прогноз можно повторять для разных запросов, а модель можно постоянно обновлять новыми данными для повышения ее производительности. В целом, эта методология предлагает эффективный и действенный способ автоматизации поиска юридической информации, связанной с преступлениями, в т.ч. уголовными, в юридических документах на арабском языке.

2.1. Подход Top2Vec

Top2Vec [9, 10] – это новая неконтролируемая техника кластеризации документов и моделирования тем, которая может обнаруживать темы в крупномасштабных наборах данных без какого-либо предварительного знания их количества. Основная

[1] Uniform manifold approximation and projection – алгоритм машинного обучения, выполняющий нелинейное снижение размерности.

[2] Hierarchical density-based spatial clustering of applications with noise – иерархическая пространственная кластеризация приложений с шумом на основе плотности.

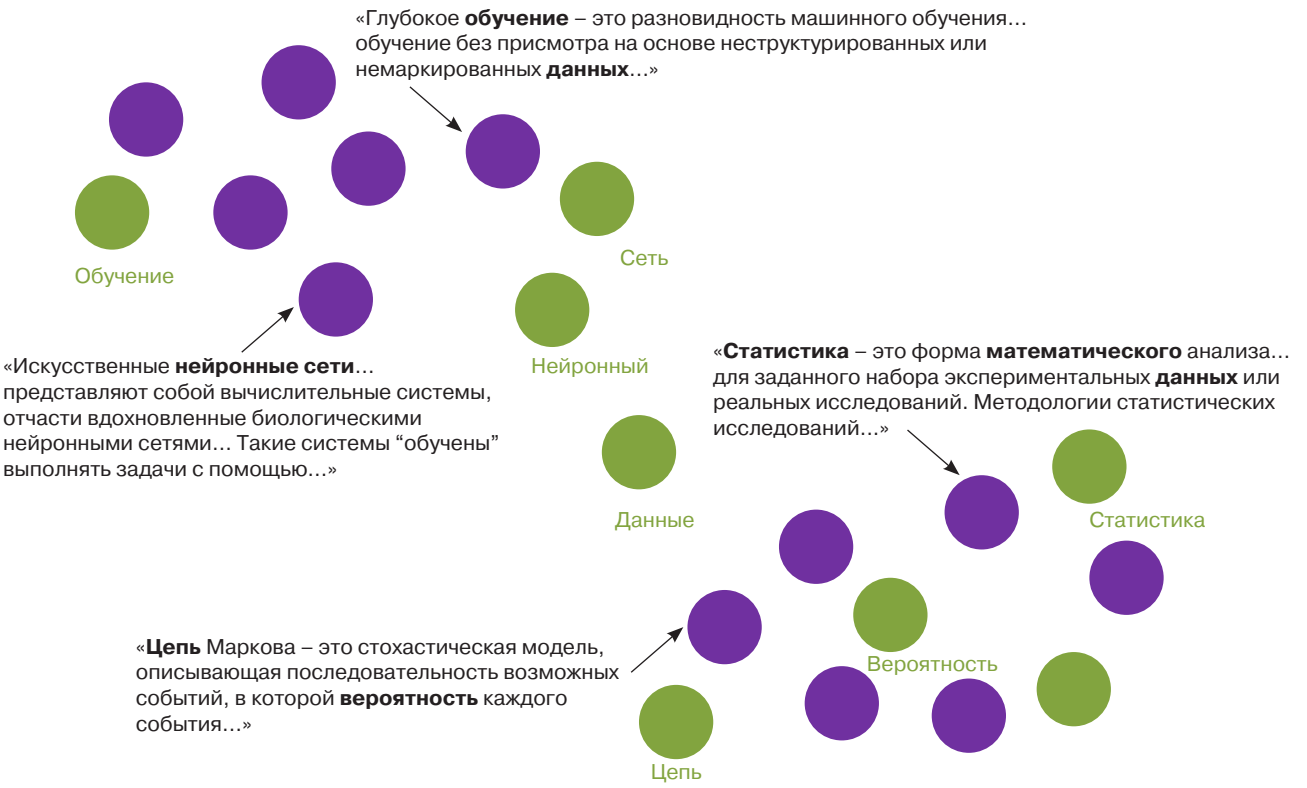

Рис. 2. Пример семантического пространства

идея Top2Vec заключается в том, чтобы встроить документы и темы в одно и то же пространство, а затем сгруппировать встроенные документы с использованием алгоритма кластеризации на основе плотности [11]. Top2Vec также может автоматически определять количество тем, а темы представлены в виде набора слов и связанного вектора. Алгоритм превосходит традиционные методы тематического моделирования, такие как скрытое распределение Дирихле и неотрицательная матричная факторизация как по качеству кластеризации, так и по масштабируемости для больших наборов данных. Кроме того, Top2Vec применялся в различных областях, включая поиск по сходству документов, визуализацию и обнаружение аномалий, где был продемонстрирован его потенциал.

На рис. 2 показан пример семантического пространства. Фиолетовые точки – это документы, а зеленые – слова. Слова ближе всего к документам, которые они лучше представляют, а похожие документы расположены близко друг к другу.

2.2. Алгоритм кластеризации HDBSCAN

Иерархическая пространственная кластеризация приложений с шумом на основе плотности (HDBSCAN) [12, 13] представляет собой алгоритм кластеризации, который может идентифицировать кластеры различной плотности и формы в многомерных пространствах. Он использует иерархический подход для построения иерархии кластеров и автоматически определяет их количество. Одним из преимуществ HDBSCAN является то, что он может обрабатывать кластеры разных размеров и форм, а также идентифицировать точки шума. Алгоритм имеет несколько гиперпараметров, которые можно настроить для оптимизации результатов кластеризации. Ниже приведены некоторые из ключевых гиперпараметров:

- Минимальный размер кластера: этот параметр устанавливает минимальное количество точек, необходимое для формирования кластера. Увеличение этого параметра приводит к увеличению кластеров и уменьшению их количества.
- Метрика: этот параметр определяет метрику расстояния, используемую для расчета сходства между точками данных. В зависимости от характеристик данных могут использоваться различные показатели.
- Метод выбора кластера: этот параметр определяет, каким образом окончательный набор кластеров выбирается из иерархии.

В процессе обучения модели Top2Vec был использован алгоритм HDBSCAN со следующими гиперпараметрами: минимальный размер кластера – 3, метрика – евклидова, метод выбора листового кластера – leaf.

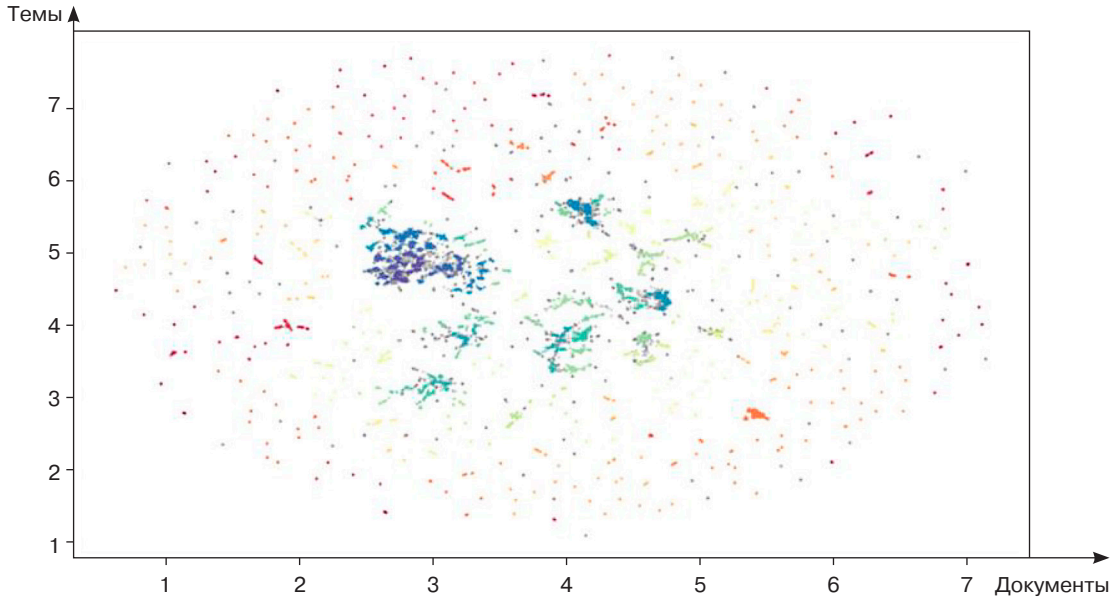

Рис. 3. Поиск плотных областей документов с помощью пространственной кластеризации приложений на основе иерархической плотности с шумом

Эти гиперпараметры выбраны на основе характеристик имеющегося набора данных и были настроены для оптимизации результаты кластеризации. Использована евклидова метрика, т.к. она является общей метрикой для измерения расстояния между точками в многомерных пространствах, и метод выбора листового кластера, поскольку он подходит для больших наборов данных и обеспечивает хороший баланс между скоростью и точностью. Минимальный размер кластера, равный 3, был выбран, чтобы гарантировать, что небольшие кластеры не будут формироваться, а шумовые точки не будут включены в кластеры.

Результирующая визуализация, показанная на рис. 3, отображает кластеры в двумерном пространстве и позволяет увидеть, как документы в каждом кластере связаны друг с другом. Этот подход представляет собой полезный инструмент для изучения и навигации по сложному ландшафту юридической информации и может помочь специалистам в области права и исследователям быстро идентифицировать соответствующие документы на основе их содержания.

2.3. Сбор данных и предварительная обработка

На этом этапе собираются и предварительно обрабатываются нормативные и юридические документы на арабском языке, касающиеся преступлений. Используется набор стандартных юридических документов, включающий различные типы юридических текстов, т.е. законы и постановления. Эти документы собраны в один файл csv, называемый *корпусом*, где каждый документ представляет собой одну запись (строку) в этом корпусе.

Каждая строка имеет несколько основных ключей, необходимых для организации данных:

1. Индекс: уникальный порядковый номер для каждого документа.
2. Идентификатор: уникальная строка для каждого документа, которая представляет собой нормативный или юридический документ, реальный идентификатор, как он упоминается в первоисточнике.
3. Название: строковое название нормативного или правового документа.
4. Резюме: строковое краткое описание содержания регламента или документа.
5. Детали: строковое полное описание и содержание документа, как оно есть в первоисточнике.

В целях контролируемого обучения модель обучается только на тех данных, которые находятся в колонке Details из корпуса. Они собираются в один список. Затем результирующий список предварительно обрабатывается, чтобы удалить стоп-слова, знаки препинания и диакритические знаки, а также выполнить определение корней и нормализацию.

Этап предварительной обработки включает в себя следующие подэтапы:

1. Токенизация: этот этап включает в себя разбиение текста на отдельные слова или токены. Это важный первый шаг в любом конвейере обработки текста, поскольку большинство алгоритмов работает с отдельными словами или токенами, а не с целыми предложениями или документами. Токенизация может выполняться с использованием различных методов: токенизация на основе пробелов, регулярных выражений и правил.

2. Удаление стоп-слов. Стоп-слова – это общеупотребительные слова, которые часто удаляют из текста, поскольку они не несут особого смысла, а их присутствие может снизить точность анализа текста. Примерами стоп-слов являются предлоги, союзы, артикли (and, the, of, in и пр.). Удаляя стоп-слова, можно уменьшить размерность данных и повысить эффективность последующих шагов обработки текста.
3. Основополагание (стемминг). Этот шаг включает в себя приведение слов к их корневой форме, также известной как основа. Цель стемминга – сгруппировать похожие слова, даже если они неидентичны. Стемминг может быть выполнен с использованием алгоритма Портера, алгоритма снежного кома и алгоритма Ланкастера.
4. Нормализация. Этот шаг включает в себя преобразование слов в стандартную форму, чтобы обеспечить согласованность и уменьшить избыточность в тексте. Примеры нормализации включают преобразование всех слов в нижний регистр, преобразование всех чисел в цифры и удаление знаков препинания. Нормализация помогает уменьшить шум в данных и повысить точность последующих шагов обработки текста.

После этого шага получаем необходимый набор данных предварительно обработанных документов, который будет использоваться для обучения модели.

2.4. Обучение модели

Обучение модели Top2Vec включает в себя несколько параметров, которые необходимо установить на основе характеристик набора данных и желаемого результата. В этой работе модель Top2Vec была обучена с использованием следующих параметров:

- Документы: корпус предварительно обработанных документов для кластеризации.
- min count: минимальное количество раз, которое должно появиться слово, чтобы быть включенным в модельный словарь.
- Модель внедрения: тип модели внедрения, используемый для создания вложений документов. В этом случае была использована модель Document-to-Vector (Doc2Vec) – неконтролируемая модель глубокого обучения, которая может изучать векторные представления документов, что позволяет эффективно вычислять сходство и моделировать темы.
- Разделить документы: логическое значение, указывающее, следует ли разбивать документы на части для обработки. В этой работе значение «раздельные документы» соответствует истине (true).
- Блокировщик документов: метод, используемый для разделения документов. В этом случае использован последовательный фрагментатор – метод разделения документов на более мелкие последовательные фрагменты фиксированной длины для облегчения обучения и обработки моделей машинного обучения.
- Длина фрагмента: максимальная длина каждого фрагмента документа. В этом случае установлена длина блока 5, т.е. каждый документ разбит на блоки по 5 предложений. Этот параметр управляет степенью детализации процесса разделения: меньшие значения приводят к меньшим фрагментам и потенциально более подробным темам, а большие значения – к более крупным фрагментам и потенциально более общим темам.
- Максимальное количество фрагментов: максимальное количество фрагментов для каждого документа. В данном случае установлено максимальное количество фрагментов равное 2, т.е. каждый документ разбивается не более чем на 2 фрагмента длиной 5 предложений. Таким образом, если документ длиннее 10 слов, он будет разделен на 2 фрагмента по 5 слов, а если документ короче 6 слов, он будет считаться одним фрагментом.
- Аргументы HDBSCAN: гиперпараметры алгоритма кластеризации HDBSCAN, минимальный размер кластера, мера расстояния и метод выбора кластера.
- Скорость: скорость для обучения модели. В работе установлено «глубокое обучение», которое является наиболее сложным и эффективным.

После установки этих параметров происходит обучение модели Top2Vec.

Оптимальные значения получаются после нескольких экспериментов с разными параметрами и сравнения результатов. Модель обучается на большом количестве юридических документов. Из-за этого и возможностей глубокого обучения процесс обучения занимает несколько часов. Результатом процесса обучения является модель, имеющая доступ к набору кластеров, каждый из которых содержит репрезентативный документ и список подобных документов. Эти кластеры можно использовать для изучения корпуса и выявления закономерностей или тем в юридических документах.

2.5. Обработка запросов

На этом этапе запрос пользователя обрабатывается так, чтобы он соответствовал предварительно обработанным документам, которые использовались для обучения модели. К запросу пользователя применяются те же методы предварительной обработки,

что и для документов. Затем предварительно обработанный запрос используется для создания набора документов-кандидатов по модели.

2.6. Поиск документов

Чтобы получить документы, связанные с запросом, с помощью Top2Vec, первым шагом является встраивание запроса в то же векторное пространство, что и документы. Это делается путем передачи запроса через ту же нейронную сеть, которая использовалась для встраивания документов. Результатом является вектор запроса, представляющий запрос в том же многомерном векторном пространстве.

Затем Top2Vec вычисляет косинусное сходство между вектором запроса и всеми векторами документа в векторном пространстве. Оценка косинусного сходства находится в диапазоне от −1 до 1, где 1 указывает на полное сходство, 0 указывает на отсутствие сходства и −1 указывает на полное несходство.

Косинусное сходство между векторами **q** и **d** можно рассчитать следующим образом:

$$\cos(\mathbf{q},\mathbf{d}) = \frac{\mathbf{q}\cdot\mathbf{d}}{(\|\mathbf{q}\|\cdot\|\mathbf{d}\|)},$$

где ||**q**|| и ||**d**|| – евклидовы нормы векторов **q** и **d**, а **q** · **d** – их скалярное произведение [14].

3. ОЦЕНКА И ОБСУЖДЕНИЕ

Предлагаемый подход к поиску документов оценивался с использованием набора реальных пользовательских запросов, связанных с преступлениями, в т.ч. уголовными, в юридических документах на арабском языке. Сначала формировался набор из 100 пользовательских запросов от юристов, работающих с такими документами. Для оценки эффективности подхода сравнивались 10 самых популярных документов по каждому запросу с набором релевантных документов, определенных экспертами в области права. Документ считается релевантным, если он содержит информацию, относящуюся к теме запроса, даже если он не содержит явного ответа на запрос. Например, документ о расследовании убийства будет считаться релевантным для запроса об убийстве, даже если он явно не отвечает на запрос. Для измерения эффективности предлагаемого подхода использованы две стандартные оценочные метрики: точность и полнота. Точность измеряет долю извлеченных релевантных документов, а полнота – долю извлеченных релевантных документов. В частности, точность и полнота определялись следующим образом [15]:

$$\text{Precision} = \frac{\text{Relevant Documents Retrieved (RDR)}}{\text{Total Documents Retrieved (TDR)}},$$

$$\text{Recall} = \frac{\text{Relevant Documents Retrieved (RDR)}}{\text{Total Relevant Documents (TRD)}},$$

где RDR – количество извлеченных релевантных документов, TDR – общее количество извлеченных документов, а TRD – общее количество релевантных документов.

Для обеспечения единой сводной меры общей производительности была проведена оценка балла F1, который является гармоническим средним значением точности и полноты. Для сравнения производительности предлагаемого подхода с базовым уровнем реализован простой подход на основе ключевых слов, который извлекает документы, содержащие хотя бы одно ключевое слово из пользовательского запроса. При сравнении использован один и тот же набор запросов и показателей оценки. Эксперименты показали, что предлагаемый подход к поиску документов превосходит подход, основанный на ключевых словах как с точки зрения точности, так и с точки зрения полноты, в частности, обеспечивает точность поиска – 87% и полноту поиска – 80% по сравнению с точностью – 66% и полнотой – 62% для подхода, основанного на ключевых словах. Оценка F1 для предлагаемого подхода составляет 0.83, а оценка F1 для подхода на основе ключевых слов – 0.63.

Эти результаты показывают, что использование методов обработки естественного языка и машинного обучения при поиске документов может значительно повысить производительность систем поиска юридической информации. При обработке пользовательских запросов и документов с использованием передовых алгоритмов предлагаемый подход позволяет извлекать соответствующие юридические документы с более высокой точностью, уменьшая нагрузку на практикующих юристов по ручному поиску в больших хранилищах документов. Тем не менее, следует отметить, что этот подход еще можно улучшить. Например, он может оказаться неэффективным при поиске соответствующих документов для узкоспециализированных юридических областей, требующих более тонкого понимания юридического языка. Тем не менее, крайне важно признать, что нынешний подход имеет возможности для совершенствования. В частности, его эффективность может быть ограничена при идентификации соответствующих документов в узкоспециализированных правовых областях, где требуется более детальное понимание юридического языка. Кроме того, потенциальные искажения, присущие обучающим данным, могут привести к искажениям в полученных результатах.

В целом, результаты проведенных экспериментов показывают, что предлагаемый подход к поиску документов является многообещающим шагом на пути к разработке более эффективных систем поиска правовой информации. Дальнейшие исследования и разработки в этой области способны привести к созданию еще более эффективных алгоритмов, которые смогут лучше поддерживать практикующих юристов в их работе.

ЗАКЛЮЧЕНИЕ

В работе предложен подход к поиску юридических документов, который использует методы обработки естественного языка, включая модель Top2Vec и алгоритм кластеризации HDBSCAN. Этот подход протестирован в сравнении с подходом, основанным на ключевых словах, и результаты показывают его превосходство с точки зрения точности и полноты. В частности, предложенный подход позволяет достичь точности поиска, равной 87% и полноты поиска, равной 80%. Это может значительно улучшить поиск юридических документов, сделав его более быстрым и точным. Однако важно отметить, что подход имеет ограничения, и в этой области есть место для будущих исследований и разработок. В целом работа подчеркивает важность использования передовых методов обработки естественного языка в юридической сфере и демонстрирует их потенциал для повышения эффективности и точности поиска юридических документов.

Вклад авторов

К.С. Жафар – провел основную часть исследования от концепции до выполнения, включая разработку методологии, реализацию алгоритма Topic-to-Vector и оценку результатов.

А.А. Мохаммад – внес вклад в обзор литературы, идентифицируя старые исследования и фоновые материалы, относящиеся к извлечению правовой информации, и помогал в написании введения.

А.Х. Исса – внес значительный вклад в написание раздела обсуждения, интерпретируя результаты в контексте существующей литературы и способствуя составлению заключения статьи.

А.В. Панов – курировал процесс исследования, предоставлял рекомендации по исследовательскому направлению и помогал в уточнении методологии и результатов.

Authors' contributions

K.S. Jafar – conducted the majority of the research from conception to execution, including the design of the methodology, implementation of the Topic-to-Vector algorithm, and evaluation of results.

A.A. Mohammad – contributed to the literature review, identifying relevant studies and background information related to legal information retrieval, and assisted in drafting the introduction.

A.H. Issa – made significant contributions to writing the discussion section, interpreting the findings within the context of existing literature, and contributed to the conclusion of the article.

A.V. Panov – supervised the research process, provided guidance on research direction, and assisted in refining the methodology and results.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Sleimi A., Sannier N., Sabetzadeh M., Briand L., Dann J. Automated extraction of semantic legal metadata using natural language processing. In: *2018 IEEE 26th International Requirements Engineering Conference* (*RE*). IEEE; 2018. P. 124–135. https://doi.org/10.1109/RE.2018.00022
2. Rogers A., Gardner M., Augenstein I. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surveys*. 2023;55(10):1–45. https://doi.org/10.1145/3560260
3. Alanazi S.S., Elfadil N., Jarajreh M., Algarni S. Question Answering Systems: A Systematic Literature Review. *International Journal of Advanced Computer Science and Applications* (*IJACSA*). 2021;12(3):359. https://doi.org/10.14569/IJACSA.2021.0120359
4. Sansone C., Sperlí G. Legal Information Retrieval systems: State-of-the-art and open issues. *Inform. Syst.* 2022;106:101967. https://doi.org/10.1016/j.is.2021.101967
5. Sartor G., Araszkiewicz M., Atkinson K., et al. Thirty years of Artificial Intelligence and Law: the second decade. *Artif. Intell. Law*. 2022;30(4):521–557. https://doi.org/10.1007/s10506-022-09326-7
6. Zhong H., Xiao C., Tu C., et al. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. 2020. *arXiv:2004.12158* [cs.CL]. https://arxiv.org/abs/2004.12158v5
7. Abu Shamma S., Ayasa A., Yahya A., et al. Information extraction from Arabic law documents. In: *2020 IEEE 14th International Conference on Application of Information and Communication Technologies* (*AICT*). IEEE; 2020;1–6. https://doi.org/10.1109/AICT50176.2020.9368577
8. Hammami E., Faiz R. Topic Modelling of Legal Texts Using Bidirectional Encoder Representations from Sentence Transformers. In: *Advances in Information Systems, Artificial Intelligence and Knowledge Management. Conference paper. International Conference on Information and Knowledge Systems.* Cham: Springer Nature Switzerland; 2023. V. 486. P. 333–343. https://doi.org/10.1007/978-3-031-51664-1_24
9. Angelov D. Top2Vec: Distributed Representations of Topics. 2020. *arXiv:2008.09470* [cs.CL]. https://arxiv.org/abs/2008.09470v1

10. Karas B., Qu S., Xu Y., Zhu Q. Experiments with LDA and Top2Vec for embedded topic discovery on social media data—A case study of cystic fibrosis. *Front. Artif. Intell*. 2022;5:948313. https://doi.org/10.3389/frai.2022.948313
11. Vianna D., de Moura E.S., da Silva A.S. A topic discovery approach for unsupervised organization of legal document collections. *Artif. Intell. Law*. 2023;Online First. https://doi.org/10.1007/s10506-023-09371-w
12. McInnes L., Healy J., Astels S. hdbscan: Hierarchical density-based clustering. *J. Open Source Softw*. 2017;2(11):205. https://doi.org/10.21105/joss.00205
13. Devlin J., Chang M.W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. *arXiv*, *preprint arXiv:1810.04805*. https://arxiv.org/abs/1810.04805v2
14. Salton G., McGill M.J. *Introduction to Modern Information Retrieval*. N.Y.: McGraw-Hill; 1983. 472 p.
15. Manning C.D., Raghavan P., Schütze H. *Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press; 2008. 492 p.

Об авторах

Жафар Камел С., аспирант, кафедра корпоративных информационных систем, Институт информационных технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: zhafar.k@edu.mirea.ru. Scopus Author ID 57552322300, https://orcid.org/0009-0004-1791-1406

Мохаммад Али А., магистрант, Факультет компьютерных наук, ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ) (109028, Россия, Москва, Покровский бульвар, д. 11). E-mail: amokhammad_1@edu.hse.ru. https://orcid.org/0009-0002-3533-568X

Исса Али Х., аспирант, кафедра автоматизированных систем управления биотехнологическими процессами, ФГБОУ ВО «Российский биотехнологический университет» (РОСБИОТЕХ) (125080, Россия, Москва, Волоколамское шоссе, д. 11). E-mail: ali.issa.rus@gmail.com. https://orcid.org/0009-0001-3699-222X

Панов Александр Владимирович, к.т.н. доцент, профессор кафедры корпоративных информационных систем, Институт информационных технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: panov_a@mirea.ru. SPIN-код РИНЦ 8571-9729, https://orcid.org/0009-0003-0310-3638

About the authors

Kamel S. Jafar, Postgraduate Student, Department of Corporate Information Systems, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: zhafar.k@edu.mirea.ru. Scopus Author ID 57552322300, https://orcid.org/0009-0004-1791-1406

Ali A. Mohammad, Master Student, Faculty of Computer Science, HSE University (11, Pokrovsky bulv., Moscow, 109028 Russia). E-mail: amokhammad_1@edu.hse.ru. https://orcid.org/0009-0002-3533-568X

Ali H. Issa, Postgraduate Student, Department of Automated Control Systems for Biotechnological Processes, BIOTECH University (11, Volokolamskoye sh., Moscow, 125080 Russia). E-mail: ali.issa.rus@gmail.com. https://orcid.org/0009-0001-3699-222X

Alexander V. Panov, Cand. Sci. (Eng.), Associate Professor, Department of Corporate Information Systems, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: panov_a@mirea.ru. RSCI SPIN-code 8571-9729, https://orcid.org/0009-0003-0310-3638