Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий - РТФ

Базовая кафедра «Аналитика больших данных и методы видеоанализа»

ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК

/ Зав. кафедрой «АБД и МВ»

_____ М.А. Медведева
(подпись) (Ф.И.О.)
02.06.2023 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)

«Разработка информационной платформы обмена данными для управления трансфером технологий»

Научный руководитель: Медведев М. А.
доцент, к.э.н
Научный руководитель: Берг Д.Б.
Профессор, д.ф.-м.н.

Студент группы РИМ-210981 Кочетов Р. В.

Екатеринбург
2023

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

РЕФЕРАТ

Тема магистерской диссертации:
«Разработка информационной платформы обмена данными для управления трансфером технологий»

Магистерская диссертация выполнена на 86 страницах, содержит 52 рисунка, 62 использованных источника. Содержание включает в себя введение, заключение, список источников и следующие главы:

1. Анализ методов машинного обучения и информационных платформ для фильтрации текстовых данных.
2. Инструментарий для разработки информационной платформы и модели машинного обучения.
3. Разработка модели машинного обучения для фильтрации текстовых данных и информационной платформы.

Ключевые слова: машинное обучение, информационная платформа, фильтрация текстовых данных.

Актуальность проекта можно объяснить тем, что он позволяет автоматизировать часть ручного труда в виде проверок созданных работ за счет внутреннего фильтра, реализованном на основе модели машинного обучения, который помогает отсеивать неподходящие по направлению статьи.

Цель магистерской диссертации – реализация платформы, представляющей из себя информационный ресурс для хранения научных работ, тематика которых связана с информационными технологиями, предварительным фильтром для которой будет модель машинного обучения, определяющая, связан ли проект с IT или его область применения имеет другое направление.

На основе данной цели были сформированы следующие задачи:

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Найти и изучить научные труды, тематика которых включает в себя использование машинного обучения и информационных платформ имеющие схожую направленность.
- В случае, если подобных работ найдено не было, изучить материалы, которые используют схожую терминологию в целях повышения осведомленности в выборе методологии.
- Реализовать модель машинного обучения, классифицирующую публикации на «целевые» (IT-область) и «другие».
- Реализовать информационную платформу, на которой пользователь сможет ознакомиться с конечным списком проектов, изучить каждый из них и выделить работы по конкретной подобласти, если таковая имеется.

Объектом исследования являются методы машинного обучения, позволяющие фильтровать данные, и методы разработки информационных платформ. Фильтрация данных подобного типа применяется в такой области, как поисковые системы, чтобы на основе запроса выдать пользователю релевантные результаты. Предмет исследования – разработка модели машинного обучения, фильтрующей текстовые данные, и информационной платформы для отображения отфильтрованных данных.

Новизна заключается в отсутствии открытой реализации подобного взаимодействия между информационной платформой и результатами работы модели машинного обучения.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 6
1 АНАЛИЗ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И ИНФОРМАЦИОННЫХ ПЛАТФОРМ ДЛЯ ФИЛЬТРАЦИИ ТЕКСТОВЫХ ДАННЫХ 10
1.1 Термины и определения 10
1.2 Обзор и анализ алгоритмов машинного обучения для фильтрации текстовых данных 13
1.3 Описание источников данных 20
1.4 Анализ применения методов машинного обучения для фильтрации данных и информационных платформ 21
1.5 Результаты и вывод по главе 37
2 ИНСТРУМЕНТАРИЙ ДЛЯ РАЗРАБОТКИ ИНФОРМАЦИОННОЙ ПЛАТФОРМЫ И МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ 38
2.1 Инструментарий для разработки модели машинного обучения 38
2.2 Инструментарий для разработки информационной платформы 39
2.3 Описание примера информационной платформы, демонстрирующей список отфильтрованных текстовых данных 44
2.4 Жизненный цикл проекта 53
2.5 Результаты и выводы к главе 56
3 РАЗРАБОТКА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ФИЛЬТРАЦИИ ТЕКСТОВЫХ ДАННЫХ И ИНФОРМАЦИОННОЙ ПЛАТФОРМЫ 57
3.1 Разработка моделей машинного обучения и их сравнение 57
3.2 Разработка серверной части информационной платформы 66
3.3 Разработка клиентской части информационной платформы 72
3.4 Результаты и выводы к главе 77
ЗАКЛЮЧЕНИЕ 78
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 80

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

ВВЕДЕНИЕ

На сегодняшний день существует множество источников в сети Интернет, которые предоставляют пользователю необходимую информацию. Обычно пользователь извлекает данные из сайтов, которые уже прошли проверку на достоверность предоставляемых фактов, так как никому не хочется оказаться неправым при использовании информации, которая была применена пользователем.

Информацию можно получить из следующих источников:

- Цифровые энциклопедии и словари.
- Официальные сайты различных государственных подразделений.
- Тематические форумы.
- Электронные библиотеки.

Все эти ресурсы объединяет то, что они являются цифровым продуктом, на разработку которого необходимы соответствующие специалисты, которые эти продукты будут в дальнейшем поддерживать, что может быть не всегда выгодно для владельца продукта с экономической точки зрения. Проблема, вероятно, не самая важная, однако, она существует.

С другой стороны, в современном мире появилась тенденция к использованию моделей машинного обучения в различных сферах общества:

- Медицинская (прогнозирование лечения больного по набору симптомов, использование машинного зрения для анализа рентген-снимков).
- Индустриальная (использование машинного зрения для поиска браков в изделии).
- Экономическая (прогнозирование курсов валют).
- Социальная (справочные чат-боты).

Данный перечень можно не ограничен приведенными примерами, что подтверждает стремление общества к использованию искусственного интеллекта (далее – ИИ) в различных деятельностях. Это объясняется тем, что

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

с помощью ИИ человек способен заменить свою деятельность либо получить дополнительный инструмент для ее осуществления. Стоит отметить, что внедрение ИИ полезно не везде. Так, например, в той же медицинской сфере модель машинного обучения не сможет полностью заменить врача с многолетним стажем, так как такую модель (так называемую экспертную систему) необходимо обучить с помощью некоторого количества врачей, обладающих тем самым опытом, без данного обучения использование модели однозначно будет невыгодным.

Однако, использование модели машинного обучения может быть использовано в более рутинной работе, например, в фильтрации чего-либо на основе ключевых слов. Такой фильтр был бы полезен при разработке электронного информационного ресурса и дальнейшем его использовании, так как пропадает необходимость вручную следить за достоверностью информации, размещенной на ресурсе.

В рамках данной работы предлагается реализовать платформу, представляющую из себя информационный ресурс для хранения научных работ, тематика которых связана с информационными технологиями (далее – IT), первоначальным фильтром для которой будет модель машинного обучения, определяющая, связан ли проект с IT или его область применения имеет другое направление. Данное описание является целью работы.

На основе цели были составлены следующие задачи:

- Найти и изучить научные труды, тематика которых включает в себя использование машинного обучения и информационных платформ имеющие схожую направленность.
- В случае, если подобных работ найдено не было, изучить материалы, которые используют схожую терминологию в целях повышения осведомленности в выборе методологии.
- Реализовать модель машинного обучения, классифицирующую публикации на «целевые» (IT-область) и «другие».

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Реализовать информационную платформу, на которой пользователь сможет ознакомиться с конечным списком проектов, изучить каждый из них и выделить работы по конкретной подобласти, если таковая имеется.

Объектом исследования являются методы машинного обучения, позволяющие фильтровать темы статей и методы разработки информационной платформы, предметом исследования – информационная платформа, взаимодействующая с данными, являющимися результатом работы модели машинного обучение в качестве фильтра.

Прежде, чем перейти к обоснованию актуальности, хотелось бы немного рассказать про популярные электронные источники информации.

В качестве примера, рассмотрим SciVal [1] и Google Академию [2]. Данные электронные ресурсы являются хранилищами научных работ и ссылок на них соответственно. Как и у всех продуктов у них есть свои недостатки: у SciVal – это не совсем очевидный интерфейс как во время создания области исследования, так и во время просмотра показателей, а также элементы выборки иногда могут полностью не соответствовать изначальным ключевым словам; у Google Академии проблема заключена в том, что он работает как обычный поисковик от данной компании, потому необходимо либо отправлять поисковой запрос с точным названием работы, либо потратить дополнительное время на изучение каждого документа на страницах выдачи результатов, что не совсем рационально.

С другой стороны, имеется форум habr.com [3], в котором каждый день происходит обзор актуальных технологий и анализ различных методологий, однако, статьи написаны больше с использованием разговорного языка и скорее субъективно (у многих статей имеются комментарии, которые не всегда могут быть полезными). Также здесь могут публиковаться статьи, тематика которых может быть совсем не связана с IT.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Проект, реализуемый в рамках диссертации, способствует решению этих проблем, объединив в себе положительные качества данных сервисов и форума.

Актуальность проекта можно объяснить тем, что он позволяет автоматизировать часть ручного труда в виде проверок созданных работ за счет внутреннего фильтра, реализованном на основе модели машинного обучения, который помогает отсеивать неподходящие по направлению статьи (в рамках диссертации выбрана тематика информационных технологий), на стороне пользователя имеется режим фильтрации, позволяющий ему выбрать интересующую подобласть, если таковая есть у основной темы.

Новизна заключается в отсутствии открытой реализации подобного взаимодействия между информационной платформой и результатами работы модели машинного обучения либо очень узкое распространение подобного взаимодействия. Проект, реализуемый в рамках данной магистерской диссертации, способствовал бы:

- Повышению уровня вовлеченности населения в изучение современных технологий и решений в области IT (как один из вариантов - внедрение реализованной и правильно функционирующей информационной платформы в сайт УрФУ в качестве дополнительного раздела).
- Упрощению создания специализированных сайтов (модель машинного обучения можно будет настроить на определение нужных тематик, не обязательно связанных с областью IT-инноваций).

Успешным выполнением данной работы является функционирующая информационная платформа, демонстрирующая список проектов, предварительно прошедших фильтрацию модели машинного обучения. Также должна присутствовать возможность более подробного изучения выбранного пользователем проекта.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

1 АНАЛИЗ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И ИНФОРМАЦИОННЫХ ПЛАТФОРМ ДЛЯ ФИЛЬТРАЦИИ ТЕКСТОВЫХ ДАННЫХ

1.1 Термины и определения

Первым и достаточно важным пунктом в любом процессе разработки является ознакомление с предметной областью, с которой предстоит дальнейшее взаимодействие. Очевиден тот факт, что существует прямая зависимость между глубиной проработки предметной области и качеством итогового продукта.

В рамках данной работы планируется взаимодействие с научными работами, тематика которых связана отраслью информационных технологий. Так как научная работа представляет из себя результаты исследовательской деятельности, то относительно IT-области исследования проводятся в виде разработки какой-либо новой технологии или методологии и изучением соответствующих источников. Успешным результатом этих процессов является создание инновации.

Инновацией называет усовершенствование старой или создание новой технологии или методологии, которая способствует повышению эффективности рабочих процессов или / и качества производимого продукта или услуги. Стоит отметить, что имеются некоторые схожие с «инновацией» по значению терминов: «изобретение», «улучшение». Термин «инновация» отличается от них уточнением, что она подразумевает создание и получение дополнительной ценности, а также связана с внедрением [4].

Инновации подразделяют на следующие типы:

- Технологические – получение нового или эффективного производства имеющегося продукта, изделия, техники, новые или усовершенствованные технологические процессы. Инновации в области организации и управления производством не относятся к технологическим.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Социальные (процессные) – процесс обновления сфер жизни человека в реорганизации социума (педагогика, система управления, благотворительность, обслуживание, организация процесса).
- Продуктовые – создание продуктов с новыми и полезными свойствами.
- Организационные – совершенствование системы управления персоналом.
- Маркетинговые – реализация новых или значительно улучшенных маркетинговых методов, охватывающих существенные изменения в дизайне и упаковке продуктов, использование новых методов продаж и презентации продуктов или услуг, их представления и продвижения на рынки сбыта, формирование новых ценовых стратегий.

В рамках сферы IT инновации, в большинстве случаев, имеют технологический тип, однако, могут косвенно быть и другим, так как способствуют улучшением соответствующей сферы. Так, например, программное обеспечение «Jira» [5] положительно влияет на сферу менеджмента, так как позволяет сотрудникам компании самостоятельно выбирать и реализовывать задачи без дополнительных уточнений с командой, так как каждый член организации видит, кто на данный момент занимается заданием и в каком оно состоянии.

Для понимания процесса разработки следует ознакомиться с инструментарием, а точнее с областью, в которой данный инструментарий применяется. Так как в рамках магистерской диссертации будет создана информационная платформа, способствующая управлению трансфером технологий и имеющая предварительный фильтр в виде модели машинного обучения, стоит раскрыть понятие следующих терминов:

- Трансфер технологий – успешное применение и / или адаптация технологии, знаний или результатов научных исследований, полученных в одной организации или сфере деятельности, для нужд другой (других) организации или сфер деятельности. У трансфера технологий имеются

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

следующие подходы: постепенное внедрение результатов в соответствующий рынок через демонстрационные мероприятия и создание прототипов; поиск технологических решений на основе потребностей конечных пользователей.

- Машинное обучение – набор методов для реализации искусственного интеллекта. Под искусственным интеллектом подразумевается особенность искусственных систем выполнять функции творческого характера, которые являются результатом мыслительной деятельности человека. Стоит отметить, что машинное обучение ставит перед собой цель не реализовать прямое решение, а обучить модель решать задачи за счет применения решений похожих задач.
- Информационная платформа – это автоматизированная информационная система особого класса, позволяющая неограниченному либо условно неограниченному кругу лиц пользоваться ее возможностями через интернет и решать свои технологические или функциональные задачи в автоматизированном режиме [7]. В свою очередь, информационной системой называют систему, взаимодействующую с цифровой информацией; информационная система хранит в себе определенную информацию, позволяет пользователям обрабатывать и искать ее.

На данный момент имеется общее определение области, в которой будет происходить реализация проекта. Следующими шагами требуется произвести анализ:

- Анализ источников научных работ.
- Анализ инструментария для реализации проекта.

Данные, полученные в результатах этих анализов, позволят выбрать наиболее удобный инструмент, источники, в которых поднимается похожая проблема, и источники, которые послужат обучающим материалом для модели машинного обучения.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

1.2 Обзор и анализ алгоритмов машинного обучения для фильтрации текстовых данных

Ранее уже говорилось, что суть машинного обучения заключается в обучении модели решать задачи за счет решений схожих. Условно, способы реализации модели машинного обучения можно разделить на классические алгоритмы и глубокое обучение. Также отдельно стоит выделить обработку естественного языка, так как это является общим направлением.

1.2.1 Классические алгоритмы

Классические алгоритмы [23] машинного обучения включают в себя:

- Линейную регрессию – алгоритм, который ищет линейную (прямую) зависимость между набором данных и строит условную линию между элементами этих данных так, чтобы каждый элемент был минимально «отдален» от нее. Расстояния могут высчитываться, например, за счет метода наименьших квадратов (минимизация суммы квадратов отклонений наблюдаемых значений зависимой переменной от значений, предсказанных моделью).
- Логистическая регрессия – алгоритм, который может быть использован для классификации (хорошо проявляет себя в бинарном варианте); для классификации используется сигмоида. Формула (1) позволяет узнать вероятность попадания в тот или иной класс.

$$P(y_i = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \qquad (1)$$

где P(yi: 1|X) – вероятность, что целевое значение i-го наблюдения y, является классом 1;

X – тренировочные данные;

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

β0 и β1 – параметры, которые необходимо обучить;

e – Эйлерово число.

- Алгоритм k-ближайших соседей – это классификационный алгоритм, который базируется на следующем принципе: рассматриваемый объект будет определен в тот класс, количество соседей которого будет наибольшим; k – количество рассматриваемых ближайших соседей.

- Алгоритм Байеса – алгоритм классификации, основывающийся на формуле условной вероятности Байеса (2).

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \qquad (2)$$

где A и B – события;

p(A) и p(B) - вероятности наступления каждого события;

p(A|B) - условная вероятность: вероятность наступления события A при условии, что произошло событие B.

- Деревья решений – это иерархические древовидные структуры, состоящие из решающих правил вида «Если ..., то ...». Правила автоматически генерируются в процессе обучения на обучающем множестве. Деревья решений могут применяться как для задачи классификации, так и для задачи регрессии. Существует алгоритм «случайного леса», где имеется множество решающих деревьев, благодаря которому при регрессии получается более усредненное значение, при классификации класс выдается тот, элементов которого будет больше.

- Бустинг – объединение одинаковых моделей, обученных последовательно, причем последующая модель должна исправлять ошибки предыдущей. Бустинг направлен на уменьшение смещения в данных, поэтому в качестве базовых алгоритмов могут браться модели с достаточно высоким смещением, например, неглубокие случайные деревья.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Бэггиннг – объединение одинаковых моделей, но обученных на различных случайных выборках из одного набора данных. Бэггинг направлен на уменьшение разброса в данных и зачастую данный прием предстает в виде алгоритма Случайного леса (Random Forest), где слабые модели – это довольно глубокие случайные деревья.
- Метод опорных векторов – метод, изначально являющийся бинарным классификатором, но имеющий возможность работать с несколькими классами. Идея метода заключается в нахождении гиперплоскости, которая максимально эффективно разделяет данные классов. Условием ее построения является максимальная удаленность от каждого класса. Элементы классов, находящиеся ближе других к такой плоскости, называются опорными векторами.

1.2.2 Нейронные сети и естественная обработка языка

Глубокое обучение предполагает создание нейронных сетей. Их отличие от моделей, основанных на классических алгоритмах, заключается в самой структуре. Искусственная нейронная сеть состоит слоев искусственных нейронов. Искусственным нейроном называют модель, основанную на человеческом нейроне и состоящую из следующих компонентов:

- Входы отвечают за поступление информации (обычно, это свойства изучаемого объекта).
- Веса – коэффициенты, позволяющие вычислять результат работы нейрона (изначально равны нулю или заданы случайным образом).
- Сумматор занимается преобразованием поступающих сигналов для их дальнейшей передачи в функцию активации.
- Функция активации определяет выходное значение нейрона в зависимости от результата на сумматоре.
- Выход – преобразованные особенности элемента.

Схема искусственного нейрона изображена на рисунке 1.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

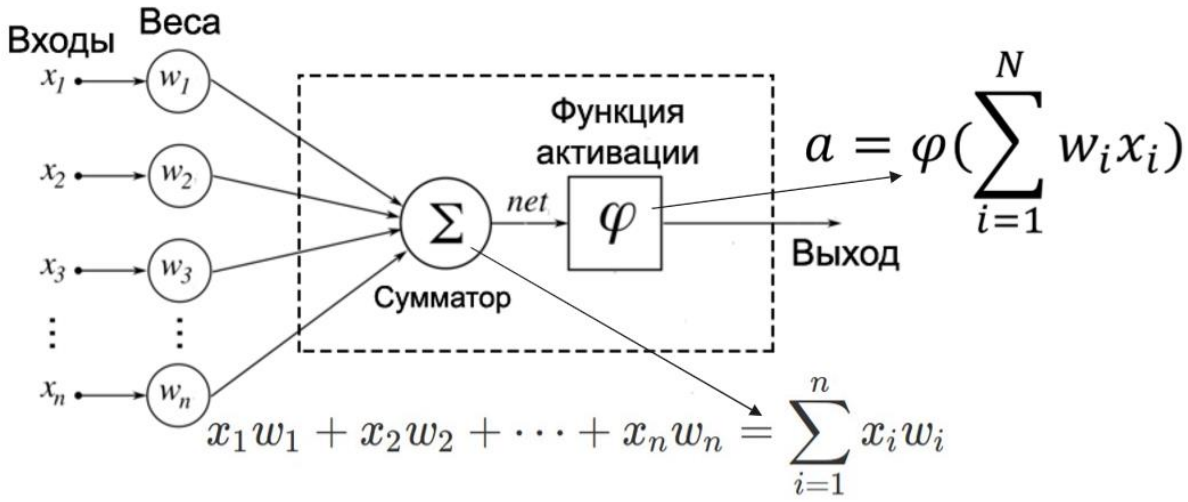


Рисунок 1 – Схема искусственного нейрона

Различают полносвязные, сверточные и рекуррентные нейронные сети:

- В полносвязных нейронных сетях каждый нейрон последовательно соединен с нейронами из соседних слоев.
- Сверточные нейронные сети хорошо справляются с обработкой изображений, так как для данных нейронных сетей нет разницы, где находится объект, когда же для полносвязных для такого же результата необходимо задать очень большое число нейронов, что вычислительно будет хуже; также сверточная нейронная сеть позволяет использовать дополнительную информацию о структуре данных: какие пиксели находятся рядом с другими.
- Рекуррентные нейронные сети благодаря своей архитектуре отлично проявляют себя при обработке естественного языка: эти сети имеют петли обратной связи в рекуррентном слое, что позволяет им сохранять информацию в «памяти» с течением времени, однако, могут возникнуть сложности с обучением стандартных рекуррентных нейронных сетей для решения задач, требующих изучения долгосрочных временных зависимостей. Это связано с тем, что что градиент функции потерь экспоненциально затухает со временем. Данную проблему решает модификация сети – LSTM: она имеет специальные регуляторы, позволяющие лучше контролировать спад градиента, и обеспечивает более долгосрочное хранение информации.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Стоит отдельно рассмотреть обработку естественного языка. Это область, объединяющая в себе сферы искусственного интеллекта и математического подхода к лингвистике и решающая задачи анализа (поиск смысла текста) и синтеза (генерация текста) языка.

Для корректной работы модели необходимо предварительно обработать текст, для этого необходимо:

1. Привести текст к единому регистру.
2. Очистить его от лишних символов (знаки пунктуации, тэги HTML и т. д.).
3. Разбить монолитный блок на атомарные составляющие, то есть провести токенизацию.
4. Разделить слова по частям речи.
5. Провести либо лемматизацию (приведение слова к его словарной форме), либо стемминг (нахождение основы слова).
6. Удалить стоп-слова (предлоги, союзы, частицы и т. д.).
7. Провести векторизацию, то есть преобразование текста в цифровое представление.

В свою очередь, векторизация может быть осуществлена несколькими подходами:

- Частотный подход – методы векторизации, заключающиеся в представлении текста в виде вектора, отображающего частоту появления слов. Алгоритм One-Hot Encoding (OHE) позволяет создать векторы N-длины, где N – количество уникальных слов, значения в векторе равны 0 или 1, обозначающие отсутствие или присутствие слова в тексте. Данная реализация применяется редко, так как она не учитывает повтор слов, потому был создан метод под названием Bag of Words, который решает данную проблему: сначала создаётся словарь всех уникальных слов, которые встречаются в тексте, текст представляется в виде вектора, где каждая составляющая соответствует частоте появления соответствующего слова. Такой подход игнорирует порядок слов и учитывает только их наличие, что может быть полезно для задач

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

классификации или кластеризации текстов. Имеется еще одна реализация – TF-IDF. Суть этого метода – вычисление оценки важности слова в документе или коллекции документов. Начало алгоритма схоже с Bag of Words, векторизация происходит таким же образом. Оценка вычисляется по формуле (3):

$$tf * idf \qquad (3)$$

где tf – показатель частоты появления слова в тексте, вычисляется по формуле (4);

idf – показатель частоты появления слова в коллекции текстов, вычисляется по формуле (5).

$$tf = \frac{n_t}{\sum_k n_k} \qquad (4)$$

где n_t – частота появления слова t в документе;

$\sum_k n_k$ – общее количество слов в документе.

$$idf = \ln \frac{|D|}{|d_i \in D \mid t \in d_i|} \qquad (5)$$

где |D| – число документов в коллекции;

$|d_i \in D \mid t \in d_i|$ – число документов из коллекции, в которых встречается слово t (при этом n_t не равно 0); логарифм нужен для сглаживания больших величин.

- Тематическое моделирование – группа методов, основанная на семантическом моделировании текстов. Данный подход пытается разделить тексты на разные кластера, но в отличии от обычной кластеризации, данный подход производит не жесткое разделение на кластера, а «мягкое», то есть каждое слово не четко относится к какому-то кластеру, а может относиться к нескольким кластерам с разной вероятностью, более правильно будет сказать,

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

что при таком подходе каждое слово может определять несколько тем (топиков), но на отнесение текста к разным темам влияет по-разному. К примеру, слово «рецепт» может с большой вероятностью говорить о том, что текст относиться к теме «Кулинария», но также может с некой меньшей долей вероятности отнести текст и к теме «Медицина». Основными методами являются LSA (латентно-семантический анализ, использует сингулярное разложение матрицы, позволяющее в дальнейшем выделить наиболее значимые признаки), PLSA (вероятностный латентно-семантический анализ, основывается на вероятностной модели; сначала создается матрицы "слово-текст", в которой каждый элемент обозначает количество вхождений данного слова в тексте, алгоритм обучения исследует данную матрицу и выделяет скрытые темы, которые объясняют наибольшую часть вариации в данных) и LDA (латентное размещение Дирихле, основан на генеративной вероятностной модели, извлекающей скрытые темы из набора текстов).

- Дистрибутивная семантика – алгоритмы, решающие задачу определения семантической близости слов на основе их распределения в большом наборе слов. Семантическое представление слова, или семантический вектор слова (он же word embedding) – это представления слова, как точки в многомерном пространстве. Данный тип вектора еще называют «плотным», так как он имеет очень компактное представление (для решения многих задач хватает вектора длиной 300 элементов). При правильном построении алгоритма и достаточно большом корпусе слов для обучения семантически близкие слова должны лежать близко в N-мерном пространстве, в котором представляется вектор слова. К примеру, слова «собака», «пес», «щенок», должны лежать ближе друг к другу, чем к слову «кот». Размерность пространства, в котором будет представлено слово, может быть различной и обычно ее подбирают под задачу.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

1.3 Описание источников данных

Для модели машинного обучения необходимо создать перечень данных, за счет которых она и обучится. Изначально рассматривались следующие варианты:

- Использование API соответствующего сервиса.
- Использование заготовленного списка научных работ.
- Самостоятельное создание подобного списка.

К сожалению, в процессе исследования данной области было установлено, что использование API сервисов требует внесения немалых материальных средств, а, например, сервис SciVal ограничил доступ к своему программному интерфейсу приложения для жителей определенного региона, то есть нет возможности даже приобрести его.

Поиск заготовленного списка научных работ на специализированных сайтах (Kaggle, Data World, UCI Machine Learning Repository) также не привел к успеху, однако, данные сервисы в целом обладают достаточным перечнем наборов данных для других целей.

Потому было решено вручную создать список научных работ. В качестве источников данных, которые понадобятся для обучения модели, были рассмотрены следующие варианты:

- SciVal – аналитическая система, предназначенная для анализа публикационной активности в библиографической базе данных Scopus [8]. Данная система позволяет не только кратко ознакомиться с содержанием работы, но и предоставляет довольно обширный инструментарий для анализа области исследования.
- Google Scholar – поисковая система от компании Google, цель которой – предоставить ссылки на научные работы, которые в какой-либо степени коррелируют с запросом пользователя.
- eLibrary [9] – крупнейшая в России электронная библиотека научных публикаций, обладающая богатыми возможностями поиска и анализа

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

научной информации. Библиотека интегрирована с Российским индексом научного цитирования (РИНЦ) - созданным по заказу Минобрнауки РФ бесплатным общедоступным инструментом измерения публикационной активности ученых и организаций.

1.4 Анализ применения методов машинного обучения для фильтрации данных и информационных платформ

С помощью сервиса SciVal было создано несколько областей исследования с целью охватить наибольший объем работ по теме диссертации (рисунок 2).

Разберем области по очереди (снизу-вверх):

1. Test research area:

- Запрос: ("transfer" AND ("information platform" OR "IT-innovations")).

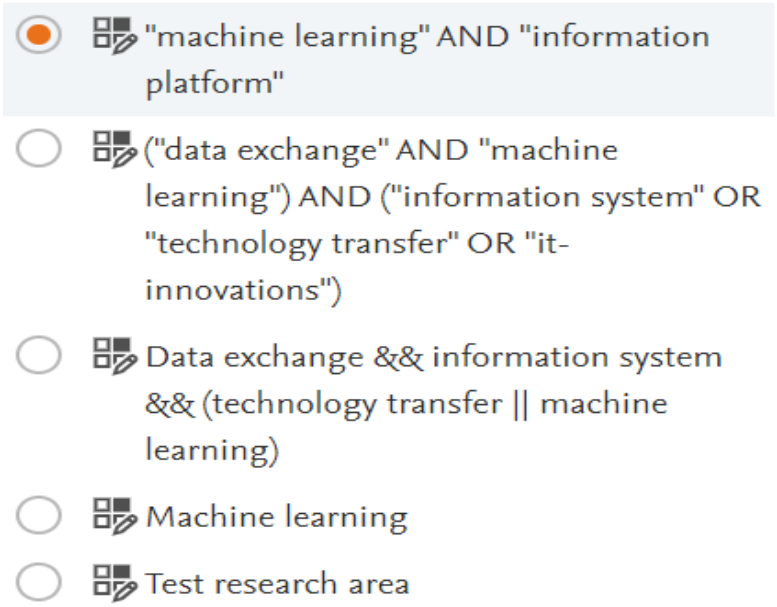


Рисунок 2 – Области исследования в SciVal

- Области: Engineering OR Computer Science.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Показатели (рисунок 3).

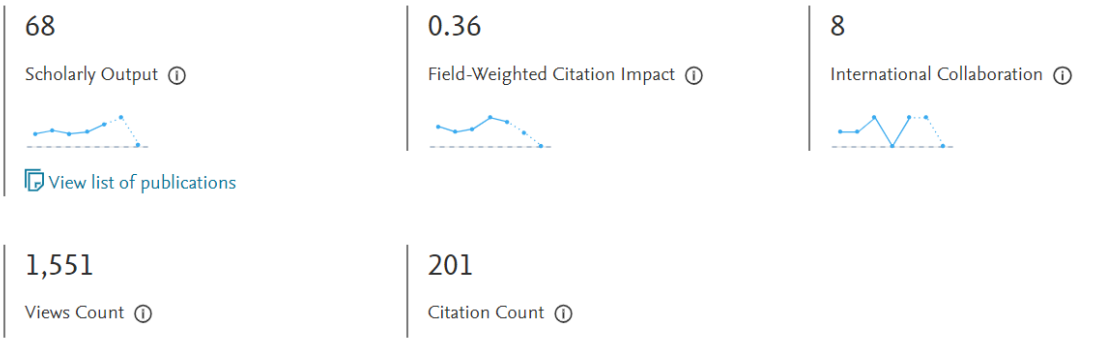


Рисунок 3 – Показатели для области исследования

- Ключевые слова (рисунок 4).

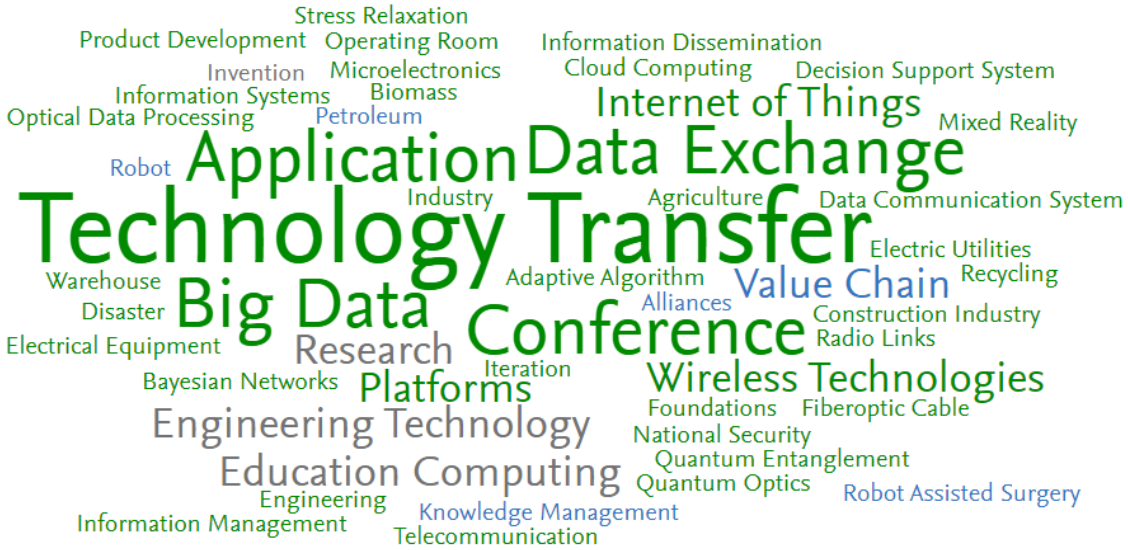


Рисунок 4 – Ключевые слова

- Выводы: данный запрос был создан для тестирования сервиса и ознакомления с платформой. Стоит отметить оптимальное количество работ, однако, их тематика больше общая (по наиболее употребляемому ключевому слову «Technology transfer» была выдача из 11 работ, из которых было изучено 3, так как тема косвенно пересекалась с темой диссертации). Показатели цитирования работ также невысокие.

2. Machine Learning:

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Запрос: (("transfer" AND "technology") AND ("IT-innovations" OR "information platform" OR "machine learning")).
- Области: Computer Science OR Engineering.
- Показатели (рисунок 5).

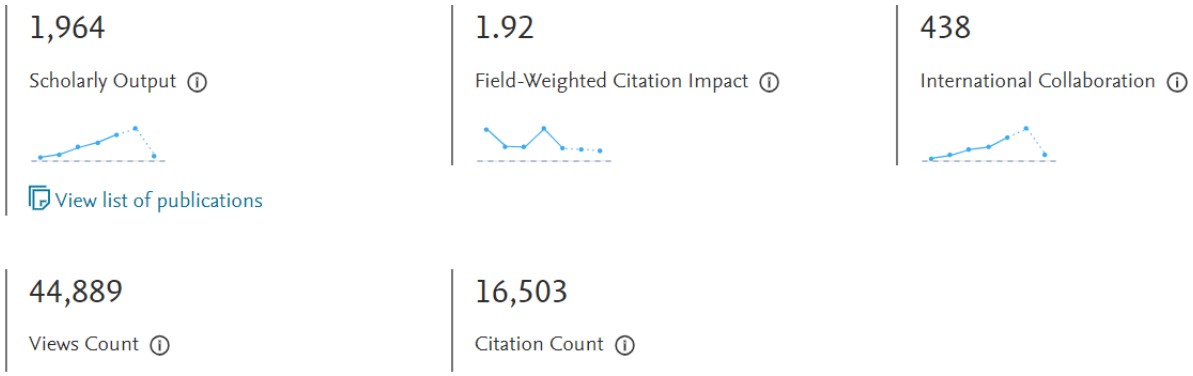


Рисунок 5 – Показатели для области исследования

- Ключевые слова (рисунок 6).

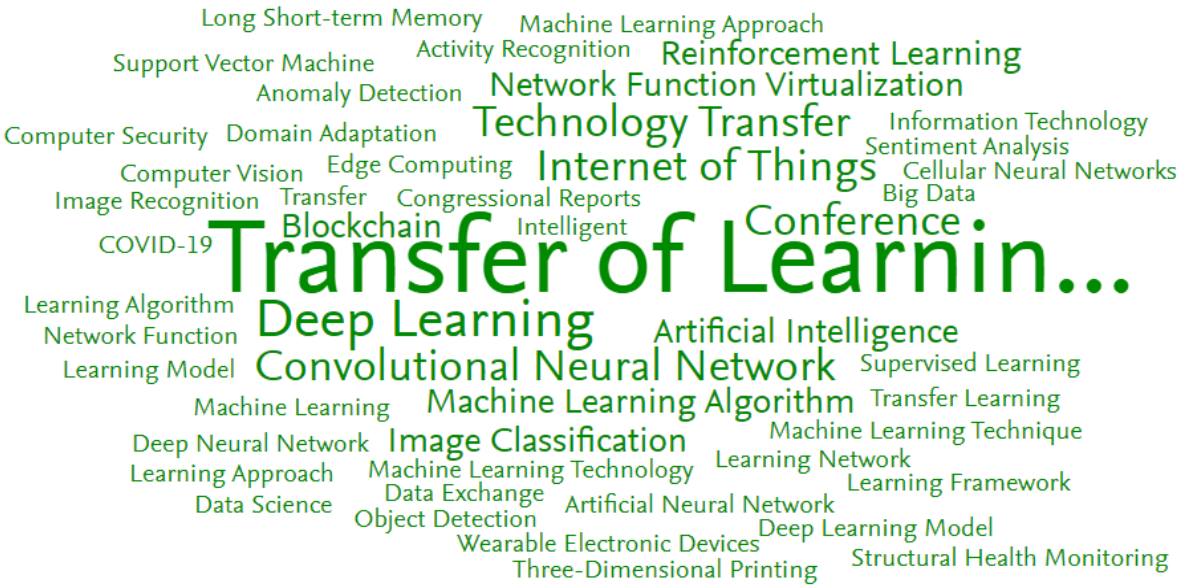


Рисунок 6 – Ключевые слова

- Выводы: из-за не совсем удачного запроса было получено огромное количество работ, большая часть которых обозревает область машинного обучения в различных его аспектах, что не совсем подходит для

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

поиска информации для реализации. Показатели цитирования высоки из-за числа научных трудов.

3. Data exchange && information system && (technology transfer || machine learning)

- Запрос: (("data exchange" AND "information system") AND ("technology transfer" OR "machine learning")).
- Области: Computer Science.
- Показатели (рисунок 7).

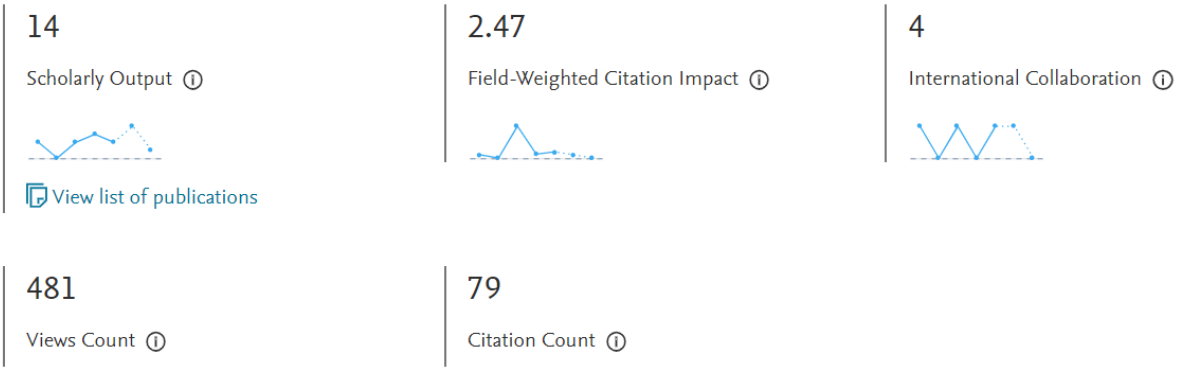


Рисунок 7 – Показатели для области исследования

- Ключевые слова (рисунок 8).

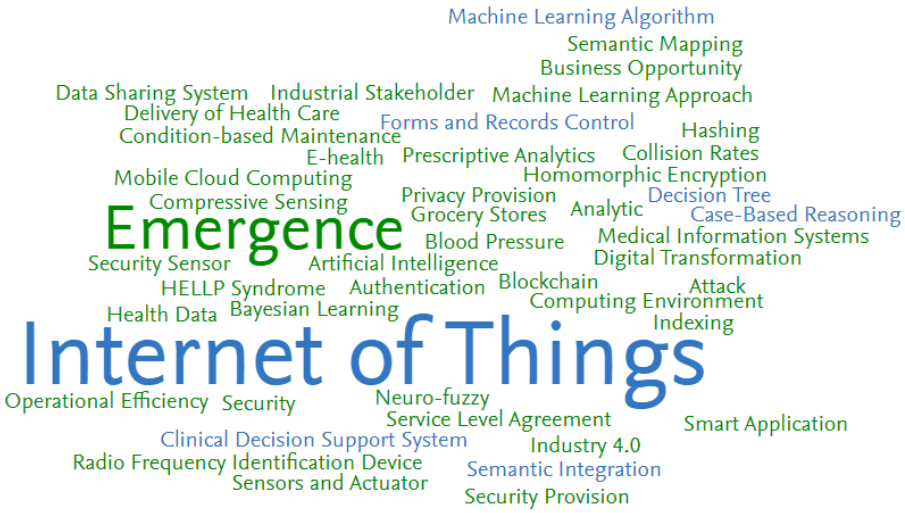


Рисунок 8 – Ключевые слова

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Выводы: в данной области исследования я решил исключить «Engineering», так как это далеко не всегда компьютерные науки. Можно заметить, что показатель взвешенного цитирования высок, однако, если рассмотреть график, то видно, что он резко перешел к нисходящей форме. Международное взаимодействие нельзя назвать активным. Интересным является факт, что наиболее распространенным ключевым словом является «Internet of Things», которое никак не было задействовано в запросе, в связи с этим большая часть публикаций связана с этим термином, но несколько работ были приняты к изучению, так как там рассматривается работа машинных моделей с семантикой.

4. ("data exchange" AND "machine learning") AND ("information system" OR "technology transfer" OR "it-innovations"):

- Запрос: аналогичен названию области.
- Данный запрос содержит те же научные работы, что и предыдущий, хотя он и был изменен.

5. "machine learning" AND "information platform":

- Запрос: аналогичен названию области.
- Области: Computer Science.
- Показатели (рисунок 9).

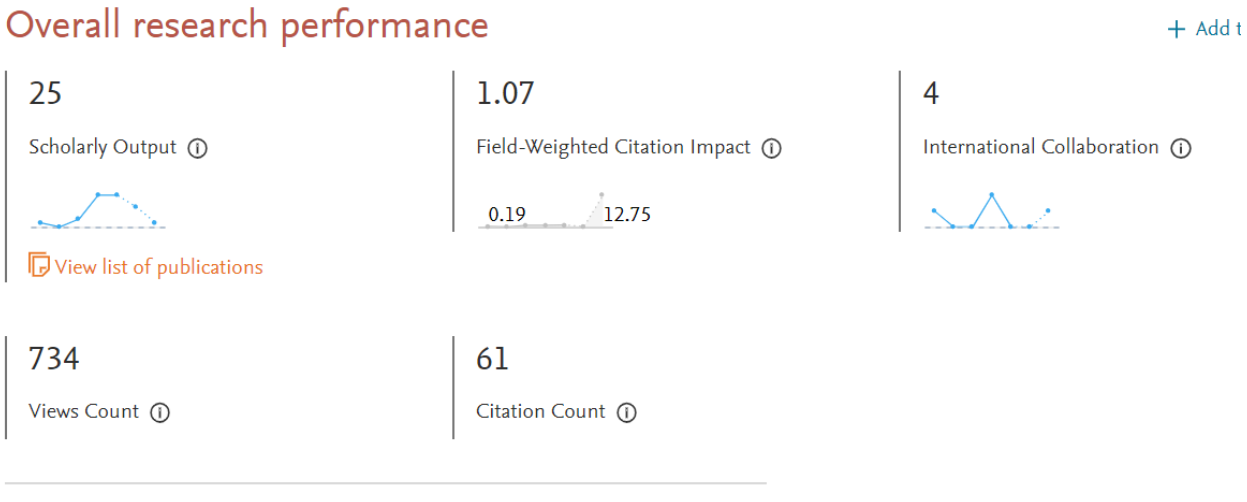


Рисунок 9 – Показатели для области исследования

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Ключевые слова (рисунок 10).

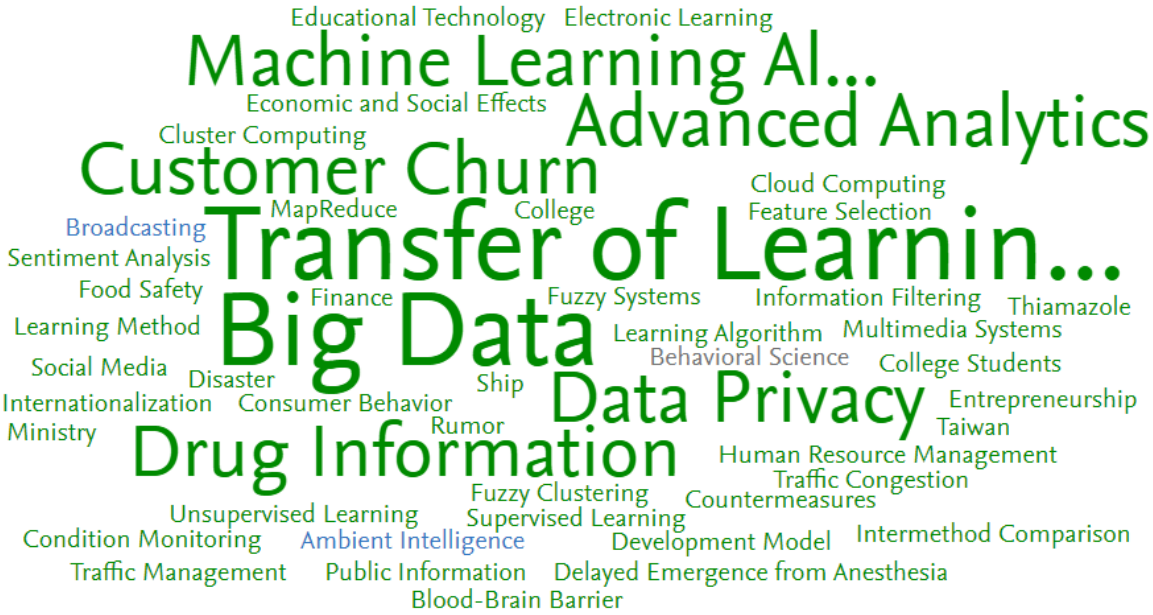


Рисунок 10 – Ключевые слова

- Выводы: данная область наиболее хорошо подходит для более тщательного изучения, так как в ней довольно много работ, связанных с созданием какой-либо информационной системы и внедрением в нее модели машинного обучения. Стоит отметить показатель взвешенного цитирования, так как в последний (2022) год он резко вырос почти в 12 раз.

Также была определена область исследования, которая наиболее точно соответствовала тематике диссертации, однако, работ по ней найдено не было (рисунок 11).

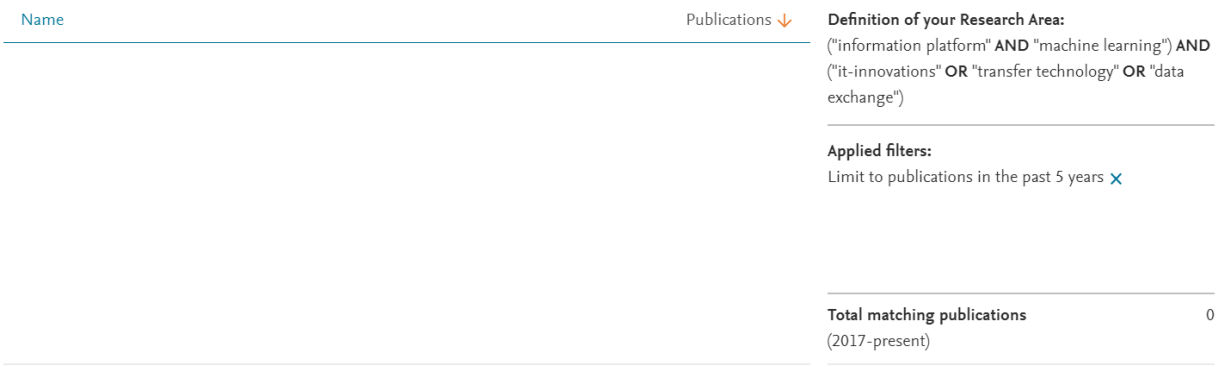


Рисунок 11 – Область исследования для тематики диссертации

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

С целью выявить особенность работ, а также определить уникальность разрабатываемого проекта был проведен обзор научной литературы из области исследования, определенной ранее.

Целью исследования, проведенного автором Liqun Ding в статье «Multimodal transport information sharing platform with mixed time window constraints based on big data» [10], было разработать платформу информационного обмена для мультимодальных (использование более одного типа транспорта) перевозок на основе больших данных с учетом смешанных временных ограничений. Актуальность данной работы заключается в том, что в условиях быстро развивающейся логистической отрасли все больше требуется эффективный обмен информацией для оптимальной организации мультимодальных перевозок. При этом такие перевозки часто имеют смешанные временные окна, которые усложняют процесс планирования и координации операций. В результате исследования была разработана платформа, которая позволяет эффективно собирать, анализировать и обмениваться данными между различными участниками логистического процесса, что позволяет оптимизировать использование ресурсов, ускорить доставку грузов и снизить затраты на транспортировку.

Авторами статьи «Sustainable Development of the Digital Economy: Global Challenges to Russia’s National Security and Prospects for Its Innovative Breakthrough» [11] ставится цель проанализировать вызовы, которые стоят перед Россией в сфере развития цифровой экономики, а также выработать рекомендации по обеспечению устойчивого развития этой области в контексте национальной безопасности. Цифровая экономика становится все более важной для развития страны, однако она также представляет ряд новых вызовов и угроз. В связи с этим необходимо провести анализ текущего состояния и определить пути достижения устойчивого развития цифровой экономики, а также обеспечения национальной безопасности. Результаты работы свидетельствуют о том, что для достижения устойчивого развития цифровой экономики необходимо сосредоточиться на улучшении

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

инновационной деятельности, создании новых технологических решений, формировании национальной экосистемы инноваций, расширении доступа к информационным технологиям, развитии электронной коммерции и обеспечении кибербезопасности.

В работе «Semantic Inference on Clinical Documents: Combining Machine Learning Algorithms With an Inference Engine for Effective Clinical Diagnosis and Treatment» [12], рассматривается возможность разработки системы для обработки клинических документов с использованием методов машинного обучения и инференции для эффективной диагностики и лечения пациентов. Актуальность статьи - эффективная обработка и анализ больших объемов клинических данных являются ключевыми задачами в медицине, поскольку это может помочь улучшить диагностику и лечение различных заболеваний. Кроме того, такая система может помочь врачам принимать более обоснованные решения на основе имеющихся данных. В результате работы был предложен ансамблевый алгоритм мультиклассификации, позволяющий одновременно запускать несколько алгоритмов классификации на основе дерева решений в одном и том же потоке данных и голосовать за наилучший результат.

Авторы статьи «Machine Learning Approach of Semantic Mapping in Polystore Health Information Systems» [13] ставят цель разработать способ оказания помощи в достижении функциональной совместимости для принятия более эффективных решений в области здравоохранения. Они считают, что важно обеспечить удобство использования и доступность данных об измерениях артериального давления в истории болезни пациента от нескольких поставщиков медицинских услуг. Актуальность данной работы связана с тем, что в полисторных информационных системах здравоохранения данные могут храниться в различных форматах и структурах, что затрудняет работу с ними. Использование семантических карт позволяет объединить эти данные и предоставить единый доступ к ним, что повышает эффективность работы с информацией в системе. Итогом исследования стал разработанный

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

алгоритм, который основан на методе «обучения без учителя», позволяющий создавать семантические карты на основе данных, хранимых в полисторных информационных системах здравоохранения. Для оценки эффективности своего подхода авторы сравнили его результаты с результатами других методов создания семантических карт, таких как методы, основанные на правилах, и методы, основанные на онтологиях. Результаты показали, что предложенный метод на основе MIL превзошел другие методы по точности и полноте при создании семантической карты из электронных медицинских записей.

Целью исследования, описываемого в статье «Integrating machine learning and open data into social Chatbot for filtering information rumor» [14], является разработка чат-бота на основе облачных вычислений и открытых данных, который сможет фильтровать недостоверную информацию. В настоящее время социальные сети и мессенджеры стали главным источником распространения дезинформации и фейковых новостей. В такой ситуации имеется необходимость в эффективных инструментах для фильтрации недостоверной информации. В рамках исследования авторы разработали чат-бота с моделью машинного обучения, основанной на скрытом семантическом индексировании (LSI) и бинарной классификации, который может автоматически анализировать текст сообщений и предоставлять пользователю соответствующую информацию о проверенных источниках.

В статье «Sentiment analysis techniques to analyze hse situational awareness at oil and gas platforms using machine learning» [15] авторы разработали модель анализа настроений персонала с использованием методов машинного обучения. Данная разработка является важной, так как она направлена на поддержание высокого уровня осведомленности персонала о ситуации на нефтегазовой платформе, что является критически важным для успешной эксплуатации нефтегазовых месторождений. В целом, это позволяет заменить ручной анализ текстового материала из отчетов персонала, который может быть чрезвычайно трудозатратным и неэффективным. В результате авторы

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

получили методологию, использующую такие методы машинного обучения, как наивный Байесовский алгоритм и рекуррентная нейронная сеть.

У данных статей достаточно высокая теоретическая база, которая может быть использована в данной работе, а авторы затрагивают различные области применения методов машинного обучения, в том числе и их использование в фильтрации текстовых данных. Так, статья [14] обозревает использование метода поиска скрытых семантических связей, который может быть использован для формирования подобластей основной тематики информационной платформы.

Однако в результате анализа библиографических источников не было найдено достаточно информации для обоснования или опровержения актуальности проекта, было решено воспользоваться патентным поиском, который позволит найти работы, которые потенциально могли бы выступить в роли прототипа проекта либо в роли сравниваемых объектов, на основе которых можно было бы выдвинуть предложение об улучшении какого-либо свойства.

Стоит отметить, что патент – это документ, выдаваемый компетентным государственным органом и удостоверяющий приоритет, авторство и исключительное право на изобретение. Это значит, что при существовании патента на информационную платформу, использующей модель машинного обучения в качестве фильтрации текстовых данных, необходимо будет предложить улучшение, которое не затрагивается в изначальном варианте, и указать, что было использовано в качестве прототипа.

В качестве источников патентов рассматривались следующие ресурсы:

- Федеральный институт промышленной собственности (ФИПС) [16].
- Яндекс.Патенты [17].
- Google Patents [18].
- eLibrary [19].

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

В качестве первоначального запроса было решено проверить, существуют ли патенты, напрямую связанные с темой данной магистерской диссертации. В ФИПС подобный запрос имеет следующий вид: Информационная платформа AND машинное обучение AND фильтрации данных. К сожалению, данный запрос не выдал каких-либо результатов (рисунок 12).

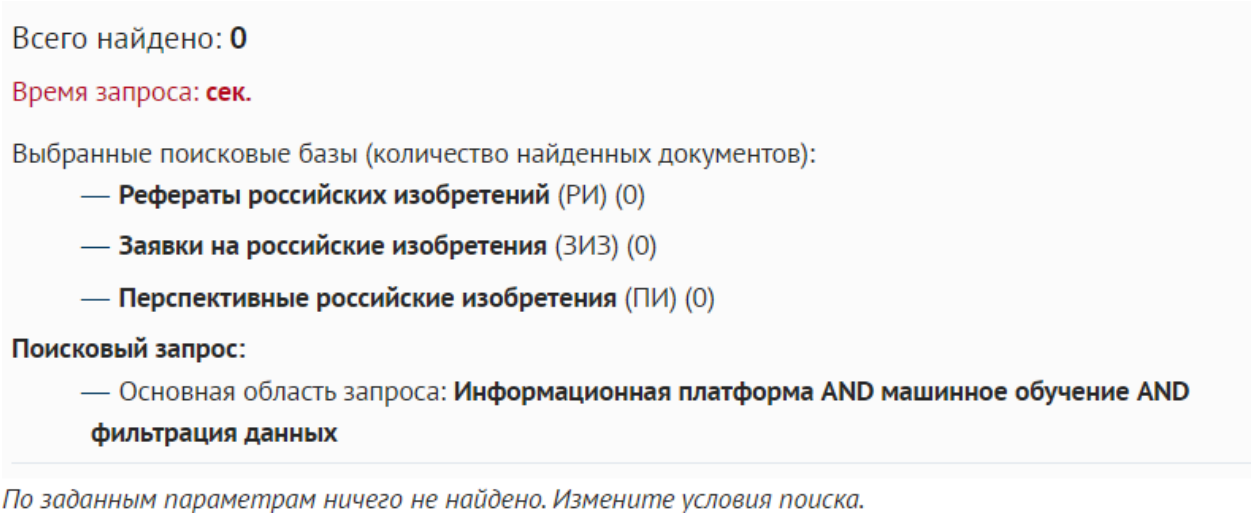


Рисунок 12 – Результат поиска

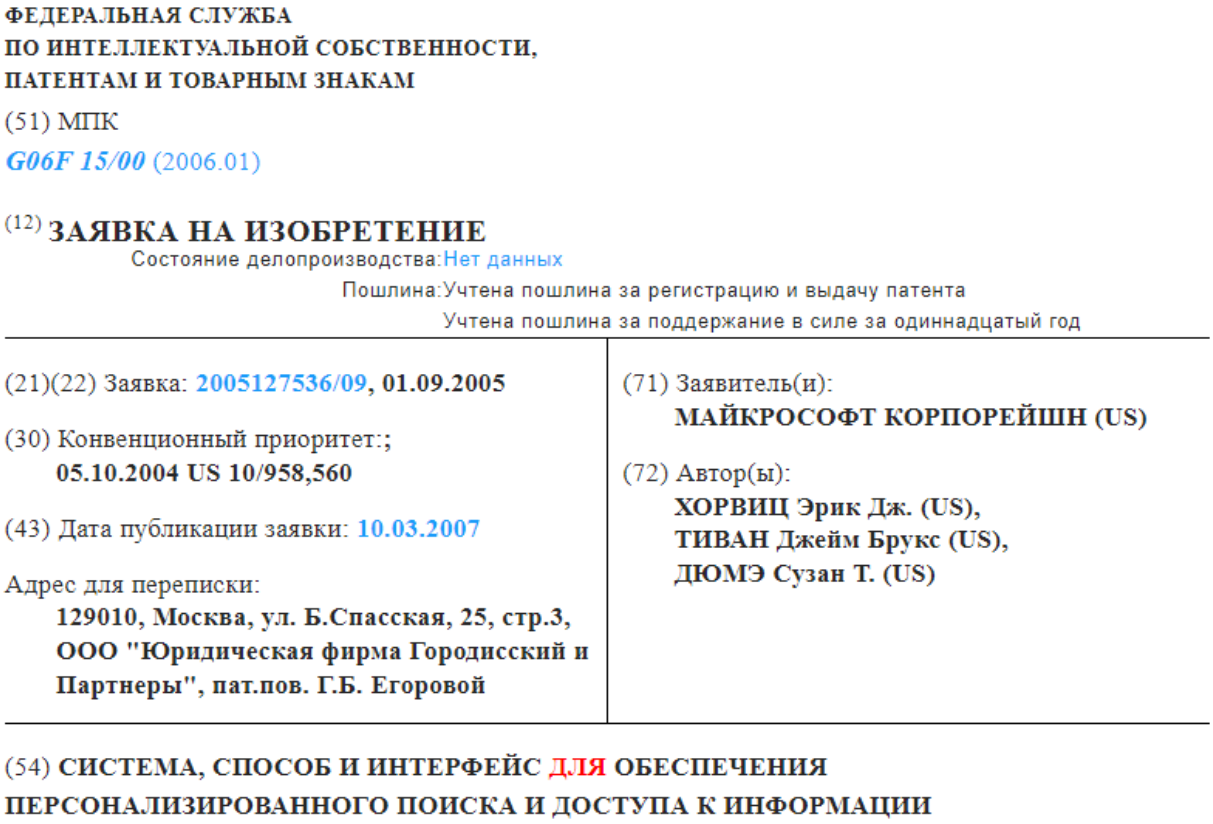
Запрос был изменен на «применение моделей машинного обучения для фильтрации данных», в результате было найдено 11 патентов, один из которых имеет схожую с магистерской диссертацией тематику: «Система, способ и интерфейс для обеспечения персонализированного поиска и доступа к информации» [20]. В рамках данного патента авторы предлагают реализовать систему, которая улучшала бы ответ на поисковый запрос пользователя. Улучшение происходило бы за счет сбора запросов пользователей и распределения их на соответствующие группы, которые отправлялись бы пользователю в качестве ответа, если его запрос находился в такой группе или пересекался по тематике с ее составляющими. Также пользовательская модель опиралась бы на каталог информации, к которой ранее обращался пользователь, и основывалась бы либо на совместной фильтрации, либо алгоритме машинного обучения, включающего в себя классификатор Байеса. У данного патенты имеются недостатки:

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Он состоит из предложений, которые можно было бы реализовать, однако, самой реализации авторами предложено не было.
- Достаточно устаревший патент (2007), с того времени появились новые способы обработки естественного языка (как пример – word2vec, который появился в 2013).

На рисунке 13 предоставлена краткая информация о патенте.



ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ,
ПАТЕНТАМ И ТОВАРНЫМ ЗНАКАМ

(51) МПК
G06F 15/00 (2006.01)

(12) **ЗАЯВКА НА ИЗОБРЕТЕНИЕ**
Состояние делопроизводства:Нет данных
Пошлина:Учтена пошлина за регистрацию и выдачу патента
Учтена пошлина за поддержание в силе за одиннадцатый год

(21)(22) Заявка: **2005127536/09, 01.09.2005**

(30) Конвенционный приоритет:;
05.10.2004 US 10/958,560

(43) Дата публикации заявки: **10.03.2007**

Адрес для переписки:
129010, Москва, ул. Б.Спасская, 25, стр.3, ООО "Юридическая фирма Городисский и Партнеры", пат.пов. Г.Б. Егоровой

(71) Заявитель(и):
МАЙКРОСОФТ КОРПОРЕЙШН (US)

(72) Автор(ы):
ХОРВИЦ Эрик Дж. (US),
ТИВАН Джейм Брукс (US),
ДЮМЭ Сузан Т. (US)

(54) **СИСТЕМА, СПОСОБ И ИНТЕРФЕЙС ДЛЯ ОБЕСПЕЧЕНИЯ ПЕРСОНАЛИЗИРОВАННОГО ПОИСКА И ДОСТУПА К ИНФОРМАЦИИ**

Рисунок 13 – Вводная информация о патенте

Далее поиск был проведен на платформе eLibrary. Прямой поисковый запрос по теме не выдал какие-либо работы (рисунок 14). Другие варианты написания тех же ключевых слов и изменение типа патента также не улучшили результат выдачи. При отдельном исследовании запросов по «машинному обучению» было получено всего 7 результатов (рисунок 15), среди которых интерес представляет патент «Связанная экспертная система» [21]. В процессе изучения работы было выявлено, что архитектура экспертной системы основана на перцептроне, что не совсем подходит под тематику реализуемого

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

проекта, так как итоговая нейросеть, основанная на перцептроне, получилась бы довольно многослойной, что увеличило бы затрачиваемые вычислительные ресурсы на ее обучение, а также она бы выглядела трудночитаемой в целом.

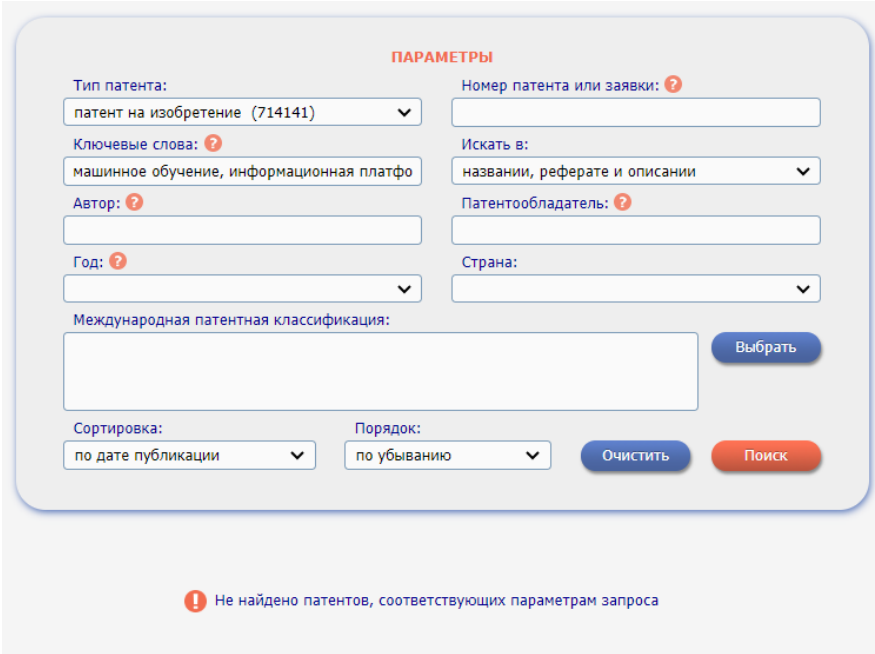


Рисунок 14 – Патентный поиск на eLibrary



Рисунок 15 – Патентный поиск и использованием ключевых слов «машинное обучение»

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Патентный поиск при помощи Яндекс.Патенты можно назвать наиболее результативным в плане количества результатов: по запросу «информационная платформа с применением машинного обучения в качестве фильтрации» было найдено 139 документов, однако, недостатком можно назвать качественную составляющую поиска, так как встречаются патенты, в которых присутствуют слова из запроса, но их тема имеет другое направление, отличное от темы данной работы (рисунок 16).

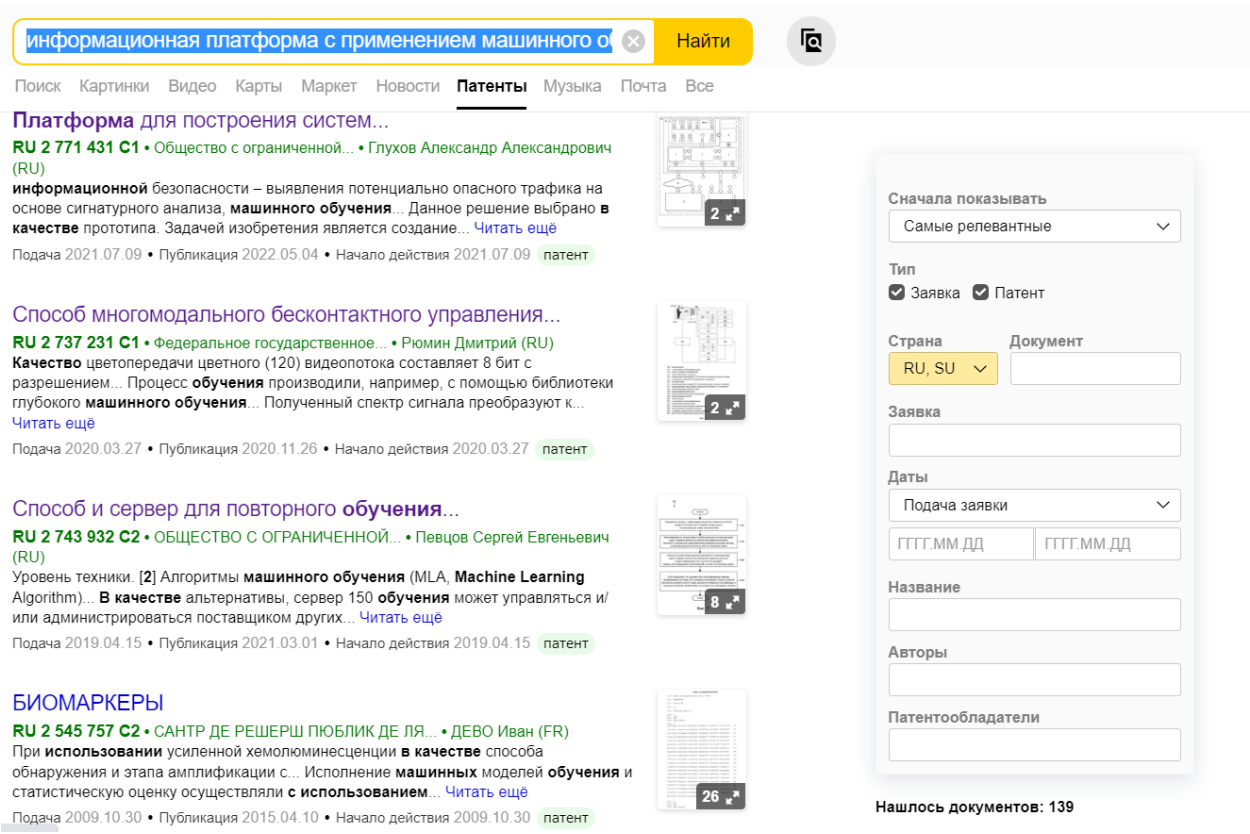


Рисунок 16 – Патентный поиск на Яндекс.Патенты

Среди данного перечня выделяется патент «Способ и сервер для повторного обучения» [22]: авторы рассказывают о повторном обучении модели Machine Learning (ML) на основе поисковых запросов с целью улучшения выдачи результатов пользователю. Суть заключается в следующем:

- Обращение сервера через сеть связи к ресурсу, содержащему множество элементов контента, каждый из которых был отправлен этому ресурсу по меньшей мере одним пользователем.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Идентификацию сервером на основе первого набора признаков по меньшей мере одного элемента контента из множества элементов контента, связанных с документом, представленным сервером поисковой системы в качестве результата поиска в ответ на поисковый запрос.
- Анализ сервером на основе набора признаков документа по меньшей мере одного элемента контента из множества элементов контента с целью определения того, был ли этот документ неверно классифицирован алгоритмом ML в ответ на поисковый запрос.
- Если определено, что документ был классифицирован неверно, формирование сервером на основе этого документа обучающего объекта, включая присвоение документу метки в виде указания на неверную классификацию, и повторное обучение сервером алгоритма ML на основе этого обучающего объекта.

В данном документе подробно описан алгоритм работы изобретения и теоретическая область (взаимодействие изобретения с текстом, с изображениями, протекающие процессы во время взаимодействия, использование метрик, предоставляемыми социальными сетями). Недостатком данной работы является ее внутренняя разработка (на момент подачи патента авторы являлись сотрудниками компании «Яндекс», данную разработку внедряли в собственную поисковую систему, на рисунке 17 изображено краткое описание).

Дополнительно был проведен поиск с применением сервиса Google Patents. При аналогичном с Яндекс.Патенты запросе в результате было получено 103 работы (рисунок 18). В целом, подходящие по теме патенты имеют ту же отрицательную сторону, что и у изобретения от Яндекса: нет возможности изучить реализацию как модели машинного обучения, фильтрующей данные и предоставляющей их информационной платформе, так и всего проекта вообще.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

(12) **ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ**

(52) СПК
G06N 20/00 *(2020.08);* ***G06F 16/906*** *(2020.08);* ***G06F 16/95*** *(2020.08)*

(21)(22) Заявка: **2019111281, 15.04.2019**

(24) Дата начала отсчета срока действия патента:
15.04.2019

Дата регистрации:
01.03.2021

Приоритет(ы):
(22) Дата подачи заявки: **15.04.2019**

(43) Дата публикации заявки: **15.10.2020** Бюл. № **29**

(45) Опубликовано: **01.03.2021** Бюл. № **7**

Адрес для переписки:
119021, Москва, ул. Льва Толстого, 16, ООО"ЯНДЕКС", патентный отдел, Малахова Надия Ахавовна

(72) Автор(ы):
Певцов Сергей Евгеньевич (RU),
Костин Михаил Юрьевич (RU),
Чигин Антон Олегович (RU),
Васильев Дмитрий Сергеевич (RU)

(73) Патентообладатель(и):
ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ «ЯНДЕКС» (RU)

(56) Список документов, цитированных в отчете о поиске: **RU 2016138553 A, 30.03.2018. RU 2664481 C1, 17.08.2018. RU 2677380 C2, 16.01.2019. US 2006/0184577 A1, 17.08.2006. US 10063434 B1, 28.08.2018. US 9324022 B2, 26.04.2016. US 2016/0155063 A1, 02.06.2016.**

(54) **СПОСОБ И СЕРВЕР ДЛЯ ПОВТОРНОГО ОБУЧЕНИЯ АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ**

Рисунок 17 – Краткое описание

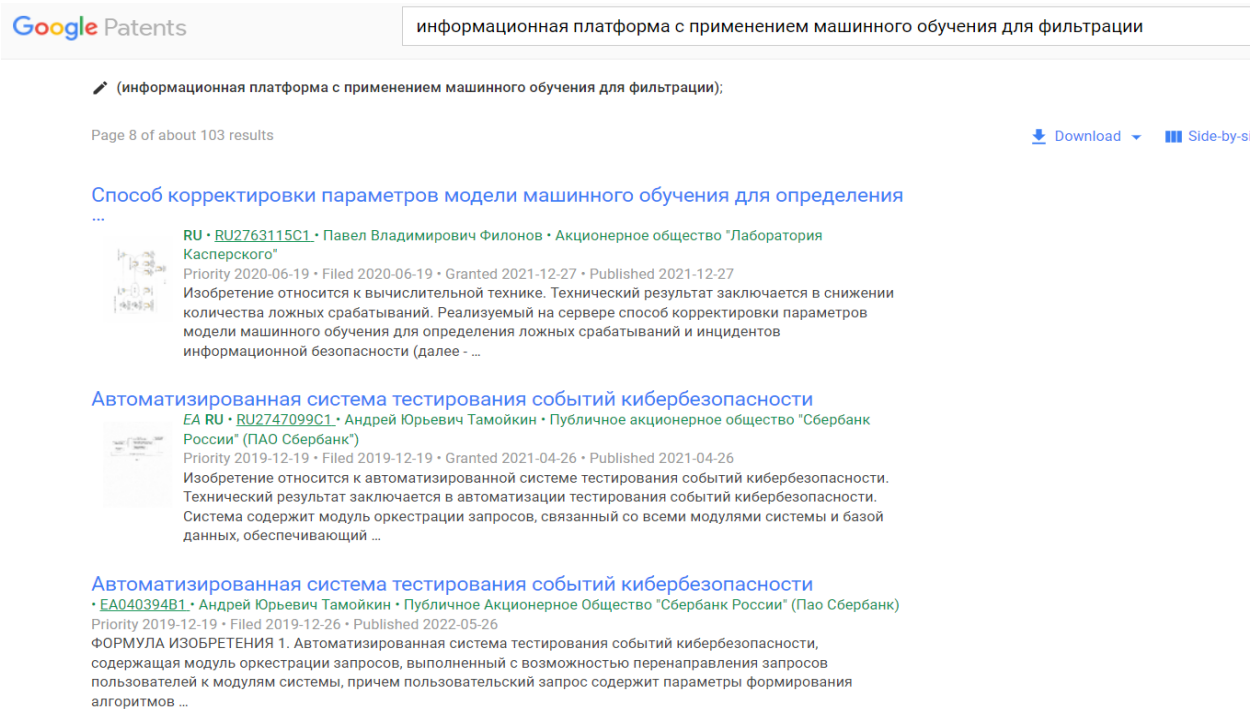


Рисунок 18 – Поиск патентов в Google Patents

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

По результатам патентного поиска можно сказать, что основным недостатком работ является их закрытость, то есть осуществление доступа к реализации доступно только определенным людям, а также давний срок подачи патента, что означает потенциальное устаревание используемых в нем технологий.

Основным отличием продукта, реализуемого в рамках проекта, будет наличие типа распространения open-source, что позволит заинтересованным в данной тематике людям изучить структуру проекта, предложить свои улучшения или просто использовать в своих проектах. В целом, разрабатываемая информационная платформа направлена на облегчение прохождения публикациями процесса модерации, так как она позволяет отфильтровать определенные экземпляры, обработка человеком которых заняла бы время на ручную проверку.

1.5 Результаты и вывод по главе

В данной главе было проведено ознакомление с терминами и определениями исследуемой области, а также были проанализированы научные статьи о применении методов машинного обучения, на основании которых была выявлена уникальность работы.

Был произведен анализ платформ, в результате которого было определено, что внедрение дополнительной фильтрации в виде модели машинного обучения будет способствовать тому, что пропадет необходимость отсеивать неподходящие по тематике работы, и проверяющему (модератору) нужно обрабатывать меньший список работ.

Были определены ресурсы, содержащие научные работы, которые послужат источником данных для обучения модели машинного обучения.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

2 ИНСТРУМЕНТАРИЙ ДЛЯ РАЗРАБОТКИ ИНФОРМАЦИОННОЙ ПЛАТФОРМЫ И МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

2.1 Инструментарий для разработки модели машинного обучения

Одной из составляющих проекта является модель машинного обучения. Для ее реализации я решил выбрать язык программирования Python, так как данный язык имеет множество библиотек для работы с машинным обучением, у которых имеется довольно понятная документация. Также немаловажным фактом является знакомство с данным инструментом, так как не придется тратить время на его изучение.

Планируется разработка следующих моделей:

- Классические алгоритмы машинного обучения из библиотеки sklearn [25].
- Использование обработки естественного языка из библиотек sklearn и genism [26].
- Модификация рекуррентной нейронной сети из библиотеки tenserflow.keras [27].

По итогу реализации моделей будет выбрана наилучшая, она и будет представлять из себя первичный фильтр, отсеивающий работы по тематике. Оценка качества будет производиться по значениям метрик accuracy, precision и recall. На рисунке 19 представлена таблица, отображающая данные метрики.

	precision	recall	f1-score	support
0	0.75	0.83	0.79	426
1	0.75	0.65	0.70	336
accuracy			0.75	762
macro avg	0.75	0.74	0.74	762
weighted avg	0.75	0.75	0.75	762

Рисунок 19 – Таблица значений classification_report

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

2.2 Инструментарий для разработки информационной платформы

Параллельно с разработкой модели машинного обучения необходимо будет реализовать саму платформу. Планируется следующее:

1. Две роли: пользователь и администратор. Различие заключается в том, что у администратора имеется функционал удаления и редактирования работы. Оба пользователя имеют возможность просматривать работы и пользоваться поиском / фильтрацией для изменения списка отображения работ.

2. СУБД, в которой будет реализована таблица для хранения работ, прошедших проверку модели машинного обучения. Выбор между обычной базой данных и системой управления базами данных пал на второй вариант, так как в случае необходимости можно будет внести изменения в данные, а также сделать резервную копию базы данных. В качестве СУБД решено было использовать PostgreSQL в связке с pgAdmin, который предоставляет удобный интерфейс управления базами данных, сокращая время через отсутствие необходимости напрямую вводить команды, связанные с модификацией базы данных.

3. Клиентская часть информационной платформы будет реализована не на чистом JavaScript, а при помощи связки ReactJS и TypeScript.

ReactJS [39] – это библиотека с декларативным подходом к разработке (происходит описание компонентов, которые браузер должен вывести, а не реализация инструкций о том, как это сделать), основанная на языке программирования JavaScript, с открытым исходным кодом, которая упрощает работу с пользовательским интерфейсом. Концепция данной библиотеки заключается в создании одностраничных приложений, то есть веб-приложений, использующих только один HTML-документ, который является оболочкой для веб-страниц. Данные подгружаются динамически, тем самым меняя структуру исходного HTML-документа, при этом обновление страницы при поступающих данных от сервера не нужно.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Одной из особенностей ReactJS является наличие такого препроцессора, как JSX, который представляет из себя объединение HTML- и JavaScript-кода. Он позволяет в одном документе создавать верстку и сразу же описывать логику, связанную с ней, а после интерпретации документа браузером вся написанная структура будет представлять из себя обычный HTML-код. Использование препроцессора не является обязательным.

Также ReactJS имеет встроенный алгоритм, позволяющий отслеживать, какой именно элемент на странице был изменен, и обновить только его, потому полная перезагрузка страницы в такой ситуации не является необходимой. Еще стоит упомянуть про наличие в ReactJS такой технологии, как Virtual DOM. Обычная объектная модель документа не приспособлена для динамической работы, так как данная структура является достаточно тяжелой, а при наличии большого количества элементов в ней их загрузка может занимать больше времени, чем это необходимо. Virtual DOM представляет из себя облегченную копию стандартного DOM, в которой проводятся необходимые действия, по завершении которых происходит их перенос на основную объектную модель документа.

К преимуществам ReactJS можно отнести возможность встроить его в уже имеющейся проект, то есть при переходе от обычного JavaScript на данную библиотеку не нужно снова создавать новый проект, ведь ReactJS может отвечать и за отдельные компоненты.

Создание компонентов в данной библиотеке может происходить двумя способами:

- Классовые компоненты.
- Функциональные компоненты.

Рекомендуется использовать функциональные компоненты, так как при создании компонентов через классы возможность их тестирования становится сложнее из-за смешивания в компоненте логики работы и визуального оформления. Используя функциональные компоненты, разработчик может сначала описать элемент, чтобы вызывать его во внешнем представлении.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Также, в отличие от компонентов класса, здесь нет необходимости использовать ключевое слово this, отвечающее за контекст вызова.

Важной концепцией работы с этой библиотекой является наличие props и state. Props – это объект, который представляет из себя подобие параметров, передающиеся компоненту, и содержащий в себе код языка программирования JavaScript. Принцип работы props следующий: объявляется компонент, которому в качестве параметра и передается этот объект, далее, используя, например, JSX, разработчик описывает HTML-код, в котором будут использоваться необходимые свойства объекта-параметра. Сами свойства задаются в вызове компонента, название свойств произвольно, задаются они подобно атрибутам в HTML.

State – JavaScript-объект, хранящий данные о компоненте, к которому привязан. Если компонент будет подвергаться каким-либо изменениям, то это отразится на его состоянии. Само состояние компонента можно изменить через методы, предоставляемые React:

- Если компонент является классовым, то используется метод setState, позволяющий изменять значение состояния компонента, свойство компонента state, в котором задаются изначальные значения состояния, а также различные методы жизненого цикла компонета (componentDidMount, componentDidUpdate и т. д.).
- Если компонент является классовым, то используются хуки – реализация методов управления состоянием из классовых компонентов для функциональных компонентов. Основными хуками являются useState (аналог state и setState), useRef (в отличие от useState не вызывает перерендера страницы при изменении значения), useEffect (совмещает в себе функционал методов жизненного цикла), useCallback (оборачивание функции таким образом, что она будет вызывать перерендер страницы только в случае изменения значений переданных зависимостей) и useMemo (аналог useCallback для переменных).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Кроме React для разработки пользовательского интерфейса существуют фреймворки Angular и Vue, однако, выбор пал на React, так как Angular имеет достаточно высокий порог входа и занимает значительно больший объем памяти по сравнению с React; от Vue было решено отказаться, так как компоненты, реализованные на нем, имеют монолитную структуру (под этим подразумевается, что логика, верстка и стили реализовываются в одном файле), что, как минимум, ухудшает читаемость кода.

TypeScript [40] – язык программирования, модифицирующий JavaScript, но не отказывающийся от его особенностей, то есть внутри файлов TypeScript также возможен ввод кода JavaScript. Одной из главных особенностей TypeScript является введение статической типизации, то есть тип переменной определяется до присваивания значения; стоит отметить, что объявление переменной по правилам JavaScript также могут быть использованы. Изначально введение типизации может показаться неудобным, однако, в будущем это поможет избежать недопонимания, с какими данными мы сейчас работаем, так как появилась возможность объявлять собственные типы.

У TypeScript есть аналог в виде Flow, однако, он менее поддерживаемый другими инструментами из экосистемы JavaScript, потому было решено его не использовать.

4. Серверную часть планируется реализовать при помощи Node.js и Express.

Node.js [41] – это программная платформа для разработки серверных приложений (возможно также использовать для разработки десктопных и мобильных приложений, но это не является основным функционалом) с открытым исходным кодом, написанная на языке программирования JavaScript. Особенностью Node.js является то, что он преобразует код, который реализован на языке JavaScript, в машинный, а это значит, что работать с данной платформой необходимо через терминал. Также стоит обратить внимание на то, что Node.js не имеет доступа к элементам объектной модели документа, так как запускается на сервере, а это значит, что объект document

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

не будет доступен для работы с данной платформой. За исключением этого момента для работы используются остальные возможности JavaScript.

Node.js был выбран, так как изучение другого языка программирования для работы с серверной частью (C#, Java, Scala, например) заняло дополнительное время, а также в комплекте с Node.js дополнительно устанавливается специальная программа Node Package Manager (npm) – менеджер пакетов, позволяющий установить на компьютер (как локально – в проект, так и глобально – на всю систему) различные архивы, являющиеся, по сути, дополнениями, расширяющими возможности разработки. Установка происходит при помощи команды npm install.

На рисунке 20 представлен процесс установки, в течение которой пользователь может наблюдать устанавливаемые файлы и ее успешный результат, в котором отображена информация о количестве задействованных пакетов, а также время, затраченное на скачивание и установку пакетов необходимого модуля.

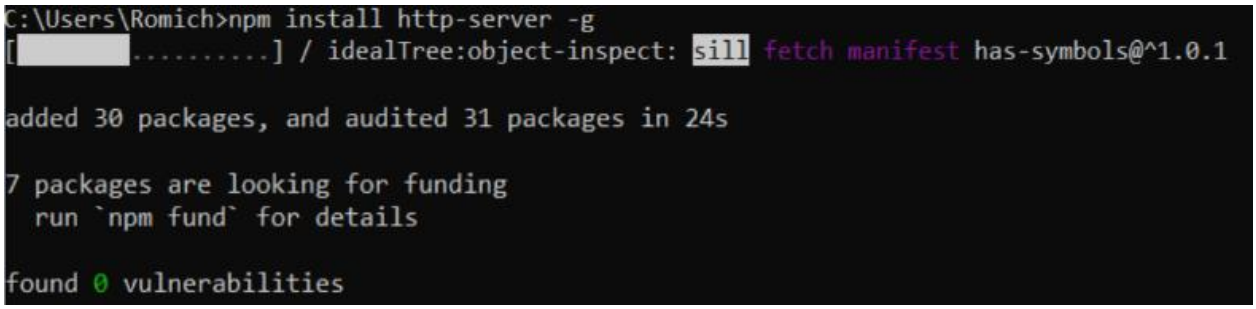


Рисунок 20 – Процесс установки пакетов через npm

Express [42] – это расширение для Node.js, являющаяся платформой для разработки веб-приложений. По сути, разрабатывать серверную часть можно и без него, однако, он включает в себя довольно удобный функционал, для реализации которого на чистом Node.js потребовалось бы сильно увеличить кодовую базу проекта.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

2.3 Описание примера информационной платформы, демонстрирующей список отфильтрованных текстовых данных

При создании любой информационной платформы следует учитывать, что необходимо реализовать следующие основные компоненты:

- База данных – цифровое хранилище информации, которая используется в процессе эксплуатации цифровой платформы. Кроме обычных баз данных существуют системы управления базами данных (далее СУБД), которые позволяют не просто хранить данные в виде файла, но и предоставляют интерфейс для взаимодействия с базой.
- Клиентская сторона информационной платформы (далее клиент) включает в себя отображение платформы в браузере, функционал, который информационная платформа должна предоставлять пользователю (интерактивность), а также корректно настроенную отправку запросов на серверную часть.
- Серверная сторона информационной платформы (далее сервер) является связующим звеном между «хранилищем» и клиентской частью, так как с клиентской части поступают запросы, обрабатываемые сервером и получающие в ответ соответствующие данные из хранилища.

Для взаимодействия между сервером и клиентом необходимо реализовать специальный программируемый интерфейс приложения или API, в котором будут описаны методы получения данных клиентом.

2.3.1 API

В общем смысле, API обозначает не только методы взаимодействия между клиентом и сервером, но затрагивает любые методы, которые могут быть использованы. В рамках реализации проекта будет использован REST

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

API [28], который использует протокол передачи данных HTTP [29]. В связи с этим могут быть использованы основные методы HTTP-запроса:

- GET – получение данных по указанному адресу и с учетом дополнительных опций, одна из которых может включать параметры к запросу.
- POST – создание нового объекта по указанному адресу и с данными, отправленными в теле запроса.
- PUT – обновление конкретного объекта по указанному адресу, обычно указывается параметр, являющейся идентификатором. Стоит обратить внимание, что данный метод именно обновляет данные указанного объекта, когда метод POST создает новый, даже при тех же свойствах, указанных в теле запроса.
- DELETE – удаление конкретного объекта по указанному адресу, обычно указывается параметр, являющейся идентификатором.

2.3.2 Клиентская часть, frontend

Любой ресурс, находящийся в пространстве Интернет, должен иметь возможность открываться с помощью браузера. Происходит открытие ресурса за счет отправки адреса ресурса через соответствующую строку, на который необходимо сделать переход: после ввода в браузере URL-адреса и нажатия определенной клавиши (обычно, этой клавишей является Enter) происходит отправка запроса на сервер на запрос на веб-сайт, идентифицированный URL-адресом, сервер формирует и выдает ответ. Важным является формат этих запросов и ответов, которые определяются протоколом HTTP – Hyper Text Transfer Protocol.

Когда происходит набор унифицированного указателя ресурса (URL) в браузере, то происходит отправка запроса с методом GET на указанный сервер, который отвечает HTTP-ответом, содержащий данные в формате HTML – Hyper Text Markup Language [30]. Браузер получает данный HTML-код и отображает его на экране. В случае, если данного URL нет, то в качестве ответа

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

возвращается ошибка с соответствующим кодом. Формат такой ошибки представлен на рисунке 21.

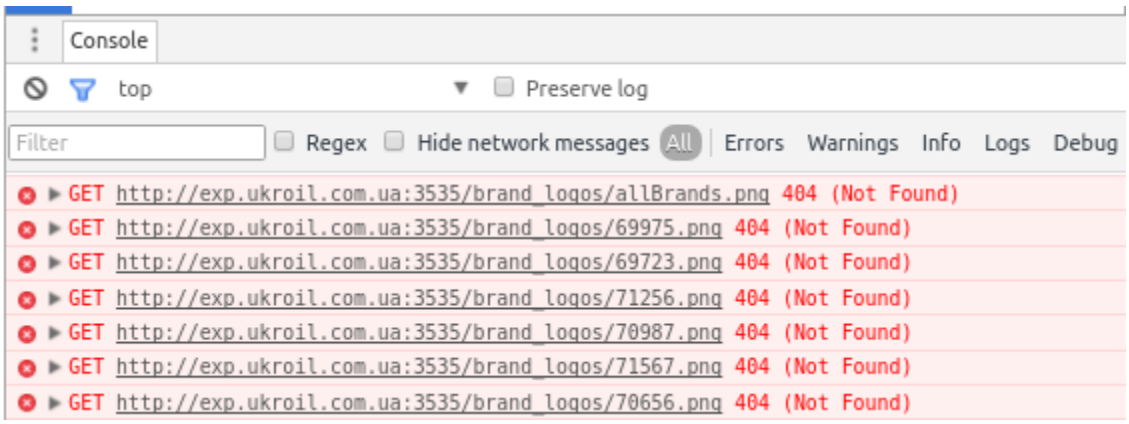


Рисунок 21 – Представление ошибки инструментах разработчика

HTML – это язык разметки текста на веб-странице, который является стандартом. Процесс отображения содержимого, написанного с помощью HTML, можно описать следующим образом:

- Происходит написание кода на языке HTML. Обычно код верстки веб-страницы состоит объявления типа страницы (от этого зависит возможность использования тех или иных функций, в современном мире распространен тип HTML 5) и тэгов, основными из которых являются тэги html (отвечает за всю страницу в целом), head и body (пишутся внутри тэга html, обычно head хранит в себе информацию, которую пользователь не видит и которая больше нужна поисковым системам, в body же происходит сама верстка).
- Далее следует интерпретация браузером написанного кода и последующее отображение уже форматированного текста, который виден клиенту.

Наполнение веб-страницы содержимым происходит за счет множества тэгов, которые имеют различное значение: это и тэги, используемые для оборачивания содержимого страницы (тэги div, section и др.), тэг ссылки (тэг

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

а), тэги перечислений (тэг маркированного списка ul и тэг нумерованного списка ol) и т.д.

Также, если необходимо, к документу может быть подключен скрипт, то есть последовательность действий, описанных на соответствующем языке программирования, с помощью тэга script. Скрипты позволяют добавить полноценное взаимодействие между пользователем и веб-страницей (отслеживание перемещения курсора мыши, нажатие кнопки, показ уведомления через определенный период времени и т. д.). Подключение может происходить как самого файла, в котором описаны необходимые действия на странице, так и написание самого кода внутри данного тэга.

Пример HTML-кода и его отображение в браузере можно увидеть на рисунке 22.

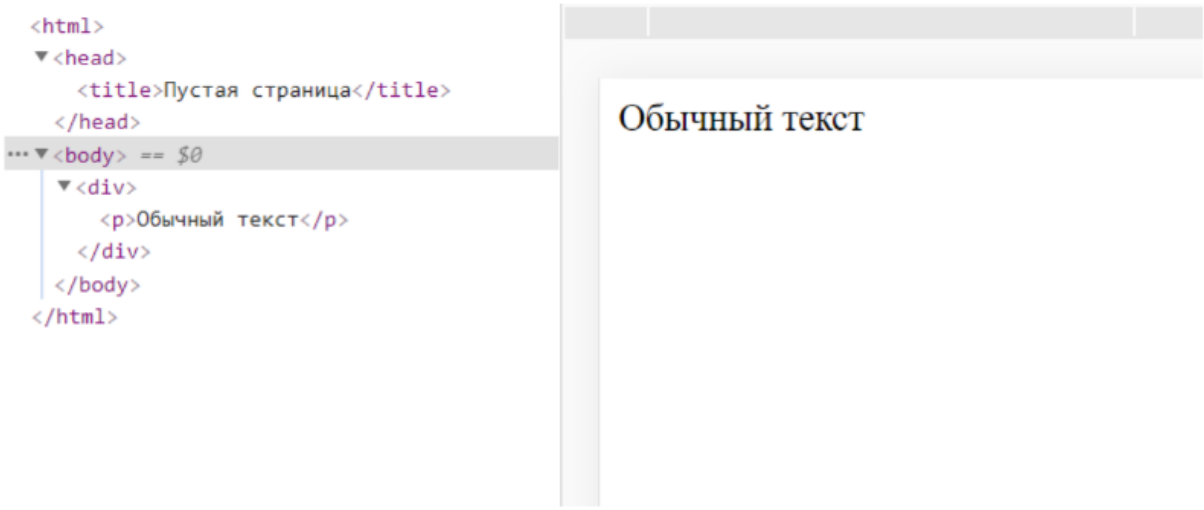


Рисунок 22 – Отображение HTML-кода на веб-странице

При получении HTML-кода происходит его парсинг браузером, то есть процесс разбора полученного кода в объектную модель документа (DOM) – специальная древовидная структура, состоящая из DOM-элементов, узлов и атрибутов и предоставляющая интерфейс для взаимодействия HTML-элементов с, например, кодом JavaScript [31]. Следующим этапом является распознавание стилей – считывание CSS-файлов тэгов style и атрибутов style

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

из DOM-дерева и их применение к элементам DOM. В конце подгружаются скрипты, написанные, как правило, на языке программирования JavaScript.

Верстка страницы в чистом виде выглядит далеко не всегда приемлемо, потому при работе с HTML практически всегда используется каскадные таблицы стилей, то есть CSS.

Если HTML необходим для создания содержимого веб-страницы, то CSS отвечает за стилизацию содержимого: расположение, цвет, фон, шрифт текста, и т. д. Описание внешнего вида элементов можно задать и в HTML-структуре, но смысл в каскадных таблицах стилей как раз заключается в том, чтобы разделить реализацию логической части веб-страницы и ее стилизацию с целью упрощения работы с версткой.

Оформление стилей с использованием файла CSS представлено на рисунке 23.



Рисунок 23 – Каскадная таблица стилей

Чтобы обратиться к элементу, используется селектор, который зависит от того, к какому элементу произошло обращение: если это обращение по тэгу, то просто пишется его название, если обращение происходит к классу элемента, то перед названием класса ставится точка, если же обращение происходит по атрибуту элемента, то в квадратных скобках пишется сам атрибут и его значение. Отдельно стоит упомянуть обращение по атрибуту id: для того, чтобы обратиться к нему, используется «#» и значение атрибута id в тэге.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Также кроме обычных селекторов существуют и специальные: псевдоклассы и псевдоэлементы. Они применяются для тех элементов, которые не отображаются в структуре веб-страницы. Различие между этими двумя понятиями заключается в том, что псевдоклассы – это селекторы для существующих элементов, а псевдоэлементы являются селекторами, которые создают новые элементы на странице. Чтобы применить стилизацию к содержимому веб-страницы, недостаточно только описать стили элементов в каскадной таблице стилей. Необходимо также подключить CSS-файл, в котором это описание и находится. Подключение происходит через HTML-структуру веб-страницы с использованием тэга link, в котором задается путь до файла стилей (если CSS-файл доступен не локально, то указывается ссылка на него в сети Интернет), тэг обычно располагается внутри тэга head.

Следующим этапом является подключение скриптов, благодаря которым открываются возможности, недоступные только с использованием HTML и CSS. Скрипты разрабатываются с помощью сценарных языков программирования. Одним из таких является язык JavaScript - объектно-ориентированный язык программирования, который работает с прототипами и имеет некоторые свойства функционального стиля. Язык был разработан в 1995 году и является реализацией спецификации ECMAScript, изначально позиционировался как язык, добавляющий интерактивность веб-страницам. С его помощью разработчики способны сделать страницу более функциональной: так при помощи данного языка программирования можно осуществить различные обработки событий, настройку форм, изменение структуры документа и его стилевых свойств. Также при помощи JavaScript реализуется взаимодействие между клиентом и сервером.

Разберемся со структурой данного языка программирования. JavaScript можно разделить на три составляющие:

- ECMAScript [32] (спецификация по языку, описывающая его составляющие: типы данных, зарезервированные слова, операторы и т. д.; по своей сути языком программирования не является).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- DOM [33] (объектное представление HTML-документа, предоставляющее доступ к методам и свойствам для изменения его структуры).
- BOM (объектная модель браузера, позволяет работать со всем необходимым окружением браузера кроме самого документа, например, информация об используемом браузере).

Подключение JavaScript-кода к странице осуществляется через тэг script, который можно объявить двумя способами:

- Напрямую написать код внутри тэга: <script>…</script>.
- Импортировать код из файла с расширением .js: <script src='здесь размещается путь до скрипта относительно данного HTML-файла'></script> (стоит отметить, что в таком случае код внутри тэга script игнорируется).

Данный язык программирования обладает следующими типами данных:

- Number – числовой тип.
- String – строковый тип.
- Boolean – логический тип.
- null – специальное значение, которое является «ничем».
- undefined – специальное значение, означающее «значение не было присвоено».
- Object – комплексный тип, который может состоять из примитивных типов (те, что объявлены ранее).
- Symbol – специальный тип, служащий уникальным идентификатором для типа Object.

Тип Object интересен тем, что включает в себя несколько подтипов:

- Array (массив).
- function (функция).
- class (является развитием функции-конструктура).
- Date (дата).
- RegExp (регулярное выражение).
- Error (ошибка).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Особенности данного языка:

- Динамическая типизация: тип данных будет определяться только тогда, когда переменной будет присваиваться ее значение.
- Гибкая работа с функциями: в JavaScript функции можно не только выполнять, но еще и возвращать функции из функций, передавать функции в качестве параметров другим функциями и присваивать функции в качестве значения переменных.
- JavaScript поддерживается всеми современным браузерами, однако, процесс разработки приложения можно осуществлять с помощью других языков (C#, Python, Java), код которых при помощи специальных библиотек преобразуется в JavaScript.
- Объектно-ориентированное программирование: методология программирования, в которой вся программа представляется в виде совокупности объектов.
- Автоматическая сборка мусора: во время работы скрипта запускается специальный сборщик неиспользуемых переменных, благодаря которому очищается память.

2.3.3 Серверная часть, backend

Кроме клиентской части в разработке сайтов используется серверная составляющая. Процесс разработки может состоять только из реализации клиентской части, но обычно это небольшие сайты (например, лендинговые страницы или страницы). Стоит отметить довольно широкое количество языков программирования, подходящих для работы с сервером.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Серверная часть же помогает расширить потенциал приложения, позволяя вместо статических данных использовать динамические, которые может получить или изменить клиент при отправке запроса на сервер, он, в свою очередь, анализирует адрес запроса, применяет соответствующий HTTP-метод и либо через SQL-запросы, либо через методы объектно-реляционного отображения (ORM) модифицирует базу данных. На рисунке 24 представлена схема взаимодействия между клиентом сервером, такой тип архитектуры называется клиент-серверным [36].

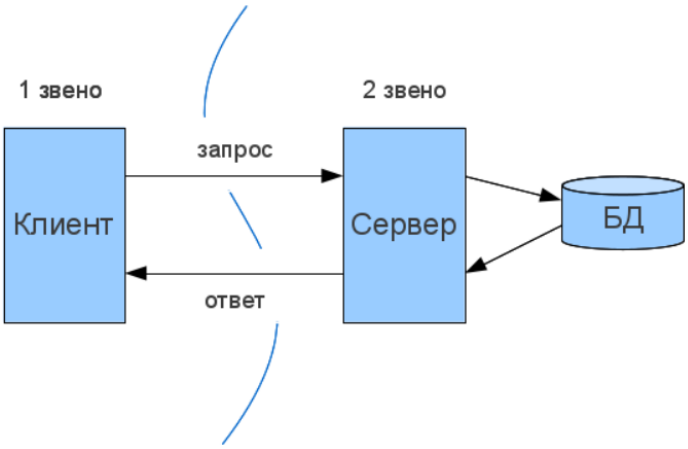


Рисунок 24 – Клиент-серверная архитектура

Если осуществлять модификацию базы данных через SQL-запросы, то необходимо описать этот запрос внутри соответствующего файла с помощью языка SQL. Операторы, с помощью которых происходит манипуляция данными:

- SELECT – выбирает данные из базы, которые удовлетворяют условиям.
- UPDATE – обновляет существующие данные.
- INSERT – добавляет новые данные.
- DELETE – удаляет данные.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

На рисунке 25 изображен запрос с оператором SELECT, реализованный с помощью языка JavaScript.

```
async  function getSearchData(tableName, allColumnName, columnNameToSearch, searchWord) {
  let arrSearchDataList;
   await connectPool
    .query(`SELECT `${allColumnName}` FROM ${tableName} WHERE `${columnNameToSearch}` LIKE "%${search
Word}%"`)
    .then((result) => {
      arrSearchDataList = result;
      // ... обработка результата запроса ...
    });

  return arrSearchDataList;
}
```

Рисунок 25 – SQL-запрос внутри языка программирования

При использовании ORM доступны специальные методы, содержащие ту же логику операторов манипуляции данными. С одной стороны, такой подход позволяет уменьшить количество печатаемого кода, а также повысить его читаемость, с другой – использование ORM ограничивает разработчика в предоставляемых методах, что лишает возможность более гибкой настройки запроса. В целом если запрос не предполагает каких-то особых условий, то использование ORM будет предпочтительнее.

2.4 Жизненный цикл проекта

Жизненный цикл проекта состоит из нескольких этапов:

1. Формирование требований – на данном этапе необходимо было сформулировать требования к конечному варианту работы. В рамках реализации данной работы были сформированы следующие требования:

   а. Проанализировать схожие реализации и выявить их недостатки.

   б. На основе проанализированных работ и статей составить задачи.

   в. Реализовать проект, способствующий решению поставленных задач (информационная платформа, использующая методы

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

машинного обучения в качестве фильтрации текстовых данных) и устраняющий недостатки аналогов.

г. Обеспечить доступ к коду итогового результата любому желающему.

2. Проектирование – разработка начальных прототипов и подготовка необходимого инструментария. В течение данного этапа составлялся дизайн клиентской составляющей информационной платформы и примерная архитектура. В целом, данный этап нельзя назвать длительным, так как итоговый проект не предполагает сложного дизайна и наличия большого функционала.

3. Разработка – процесс реализации проекта по сформированным ранее требованиям. Данный этап включает в себя три подэтапа:

а. Разработка модели машинного обучения, которая будет использоваться в качестве фильтра текстовых данных (в рамках магистерской диссертации текстовыми данными являются названия научных работ). В процессе разработки планируется реализовать несколько моделей с целью выбора наилучшей. Результатом работы модели служит исходный перечень данных с новым свойством, который хранит значение, определенное моделью и являющееся показателем прохождения фильтрации (1 – фильтрация пройдена, 0 – элемент из исходного набора данных фильтрацию не прошел).

б. Разработка серверной составляющей проекта состоит из разработки сервера для информационной платформы и реализации взаимодействия между базой данных и обработанными данными, полученными в результате работы машинной модели. Для сервера информационной платформы необходимо будет создать связь с базой данных, а также планируется использовать три метода HTTP-запросов: получение списка работ (GET), создание (POST) и удаление работы (DELETE). Взаимодействие между базой данных и моделью машинного обучения будет реализовано в области кода модели.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

в. Разработка клиентской части информационной платформы предполагает создание пользовательского интерфейса, основанного на ранее созданном дизайне, реализацию функций удаления, создания работ и получения их списка из базы данных, а также налаживание взаимодействия с серверной составляющей информационной платформы. Стоит отметить, что платформа будет использовать ролевую модель, определяющую допустимый функционал для пользователя.

4. Тестирование – проверка результатов работы происходит следующим образом: создается пример работы с соответствующим значением целевой метки, данная работа выбирается из базы данных и передается модели машинного обучения для проверки на соответствие тематике. В результате проверки работе присваивается свойство со значением, определяющим успешность проверки, в случае успеха которой работа будет отображаться на клиентской стороне информационной платформы.

5. Внедрение – в качестве заинтересованной во внедрении стороны выступает кафедра «Аналитика больших данных и методы видеоанализа», по завершении предыдущих этапов происходит составление плана по внедрению программного продукта в инфраструктуру, само внедрение будет происходить вне рамок магистерской диссертации.

6. Эксплуатация – предполагается, что на этапе внедрения будет обсужден вопрос о предоставлении удаленного сервера, автоматически запускающего процесс проверки работы на соответствие основной тематики информационной платформы. В целом, во время эксплуатации происходит процесс использования платформы пользователями и параллельное ее улучшение архитектуры и исправления проблем, найденных пользователями на данном этапе. Процесс обучения использованию не займет много времени. Важной деталью является наличие возможности обучить модель обрабатывать другие данные.

Вывод из эксплуатации – данный этап подразумевает либо отсутствие необходимости в применении разработки, либо ее замена на более выгодный

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

вариант. Относительно программного обеспечения, разработанного в рамках магистерской диссертации, вывод из эксплуатации будет происходить по второму сценарию, так как одной из главных особенностей проекта является частичная замена деятельности человека, что означает его постоянную актуальность.

2.5 Результаты и выводы к главе

В рамках второй главы был определен инструментарий, который будет использован для реализации проекта, описаны его свойства, которые могут быть использованы при разработке, а также приведены основания, почему был выбран именно данный инструмент.

Была дана подробная характеристика машинному обучению, описаны виды классических алгоритмов машинного обучения, затронуты виды нейронных сетей.

Было дано описание информационной платформе, ее структуре, определены такие понятия, как API, клиентская сторона информационной платформы и серверная сторона информационной платформы.

Следует отметить, что в данной главе приведены далеко не все возможности используемого инструментария, целью было дать описание тех инструментов и методов, которые будут использованы в процессе реализации.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

3 РАЗРАБОТКА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ФИЛЬТРАЦИИ ТЕКСТОВЫХ ДАННЫХ И ИНФОРМАЦИОННОЙ ПЛАТФОРМЫ

3.1 Разработка моделей машинного обучения и их сравнение

Реализация модели машинного обучения происходила с помощью языка программирования Python и его библиотек:

- Psycopg2 [43] – библиотека, необходимая для взаимодействия с СУБД PostgreSQL. С ее помощью будет происходить изъятие имеющихся в базе данных работ и их изменение.
- pandas [44] – библиотека для работы с наборами данных (dataset).
- sklearn – библиотека, содержащая инструментарий для работы с алгоритмами машинного обучения. В данной работе понадобятся метрики из этой библиотеки для сравнения алгоритмов.
- gensim – библиотека, реализующая методы обработки естественного языка.
- nltk [45] – библиотека, содержащая инструментарий для предварительной обработки текста.
- tenserflow.keras – библиотека, позволяющая реализовать нейронную сеть. В рамках данной работы будут реализовываться RNN и LSTM.

Для обучения модели необходимо использовать данные для тренировки. Их можно получить при доступе к API электронных научных библиотек (обычно этот доступ является платным, цена за который может варьироваться от 30 до 40 тысяч рублей), при нахождении общедоступного набора данных (очень важен набор этих данных и их свойства, так как от этого очень сильно зависит качество обученной модели) или при ручном составлении (способ не является идеальным, так как требует большего количества времени по

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

сравнению с другими методами, однако, позволяет самостоятельно сформировать таблицу, в которой можно определить как свойства данных, так и сами данные). В связи с недостаточной материальной составляющей и отсутствие общедоступного набора необходимых данных, было решено создать перечень для обучения вручную.

На рисунке 26 изображен данный набор данных. Основным столбцом является name, так как на основе имени работы будет происходить обучение модели.

id	name	is_IT_area	subtype
1	МЕТОДИКА КОМПЛЕКСНОЙ ОЦЕНКИ ЭФФЕКТИВНОСТИ ГРАФИЧЕСКИХ ПОЛЬЗОВАТЕЛЬСКИХ ИНТЕРФЕЙСОВ ИНФОРМАЦИОННЫХ СИСТЕМ МЧС РОССИИ	1	UI
2	ОПРЕДЕЛЕНИЕ МНОГОКРИТЕРИАЛЬНОГО ПОКАЗАТЕЛЯ КАЧЕСТВА ГРАФИЧЕСКОГО ИНТЕРФЕЙСА ПРОГРАММНО-АППАРАТНОГО КОМПЛЕКСА СВЯЗИ	1	UI
3	ТЕРМИНОЛОГИЧЕСКИЙ БАЗИС ОЦЕНКИ ПОЛЬЗОВАТЕЛЬСКИХ ИНТЕРФЕЙСОВ: ОБЗОР СТАНДАРТОВ	1	UI
4	ОЦЕНКА КАЧЕСТВА ПРОЕКТИРОВАНИЯ ПОЛЬЗОВАТЕЛЬСКИХ ИНТЕРФЕЙСОВ НОВОГО ПОКОЛЕНИЯ	1	UI
5	ОСОБЕННОСТИ ПОЛЬЗОВАТЕЛЬСКОГО ИНТЕРФЕЙСА ЦИФРОВЫХ СЕРВИСОВ ГОСУДАРСТВЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ, ПРЕДНАЗНАЧЕННЫХ ДЛЯ ГОСУДАРСТВЕННЫХ СЛУЖАЩИХ	1	UI
6	ОЦЕНКА ПОЛЬЗОВАТЕЛЬСКОГО ИНТЕРФЕЙСА ИНФС	1	UI
7	РЕШЕНИЕ ПРОБЛЕМЫ НЕСТАБИЛЬНОСТИ АВТОМАТ	1	UI
8	РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ДЛ	1	UI
9	ПРОЦЕСС АВТОМАТИЗАЦИИ МНОГОКРИТЕРИАЛЬНС	1	UI
10	МОДЕЛЬ МНОГОКРИТЕРИАЛЬНОЙ ОЦЕНКИ КАЧЕСТВА ПОЛЬЗОВАТЕЛЬСКОГО ИНТЕРФЕЙСА	1	UI

Рисунок 26 – Сформированный набор данных

Данный набор данных при помощи библиотеки pandas импортируется в рабочее пространство, на его основе создается DataFrame [46] – специальная таблица, являющаяся, по сути, двумерным массивом.

Таблица необходима для хранения изначальных данных и добавления новых, одними из которых являются токены – предварительно обработанный и разделенный на отдельные составляющие. При обработке текста создается

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

дополнительный столбец, основанный на полученных токенах и представляющий из себя их перечисление. Код, осуществляющий обработку текста, изображен на рисунке 27.

```
# убрать ссылки
def remove_URL(text):
    url = re.compile(r'https?://\S+|www\.\S+')
    return url.sub(r'',text)

# убрать HTML разметку
def remove_html(text):
    html=re.compile(r'<.*?>')
    return html.sub(r'',text)

# убрать эмодзи
def remove_emoji(text):
    emoji_pattern = re.compile("["
                           u"\U0001F600-\U0001F64F"  # emoticons
                           u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                           u"\U0001F680-\U0001F6FF"  # transport & map symbols
                           u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                           u"\U00002702-\U000027B0"
                           u"\U000024C2-\U0001F251"
                           "]+", flags=re.UNICODE)
    return emoji_pattern.sub(r'', text)


stop_words = set(stopwords.words('russian'))
def mytokenize(text, normalize=None, stop_words=stop_words):
    text = remove_html(text)
    text = remove_URL(text)
    text = remove_emoji(text)
    text = text.lower()
    text = word_tokenize(text, 'russian')
    text = [word for word in text if word.isalpha()]
    text = [word for word in text if not word in stop_words]
    if normalize == "s":
        porter = PorterStemmer()
        text = [porter.stem(word) for word in text]
    if normalize == "l":
        lemmatizer = WordNetLemmatizer()
        text = [lemmatizer.lemmatize(word) for word in text]
    return text

from functools import partial

mytokenize_s = partial(mytokenize, normalize='s')
mytokenize_l = partial(mytokenize, normalize='l')

data1['tokens'] = data1['name'].apply(mytokenize_s)
```

Рисунок 27 – Код для обработки текста

По завершении обработки данные разделяются на тренировочные (train) и тестовые (test) при помощи функции из sklearn train_test_split. Тренировочные данные нужны для обучения модели, а тестовые служат для проверки обучения. Очень важно следить за этими показателями так, чтобы

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

train не был слишком высоким (возможно переобучение – результат работы модели, который имеет высокие показатели метрик качества на данном наборе данных, но показывающий худший результат с другим) и слишком низким (неудачно подобраны параметры).

Для выбора модели, которая будет выступать в качестве предварительного фильтра научных работ, были рассмотрены следующие реализации:

1. Doc2bow – реализация «мешка слов», которая, в целом, справляется с задачей классификации, но успешным этот результат не назвать (рисунок 28). Кроме проблем с определением работ с неподходящей тематикой и низкой точностью (55%) данный метод ограничен изначальным набором слов, что делает его плохим вариантом для применения на долгосрочный период.

```
model = SVC(C=2, gamma=0.4).fit(xtrain, ytrain)
y_val_train = model.predict(xtest)
print(classification_report(ytest, y_val_train))
```

```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         9
           1       0.55      1.00      0.71        11

    accuracy                           0.55        20
   macro avg       0.28      0.50      0.35        20
weighted avg       0.30      0.55      0.39        20
```

Рисунок 28 – Метрики реализации doc2bow на тестовых данных

2. TF-IDF – данная реализация одноименного метода интересна тем, как высчитывается важность слова в текстах (в данном случае – научных работах). По сравнению с предыдущим методом повышается точность определения причисления работы к тематике (с 55% до 80%), однако, остается проблема ограничения слов, потому при использовании данного метода возникает необходимость постоянного обучения модели для поддержания ее в актуальном состоянии. Показатели метрик реализации представлены на рисунке 29.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

```
model = SVC(C=2, gamma=0.4).fit(xtrain, ytrain)
y_val_pred = model.predict(xtest)
print(classification_report(ytest, y_val_pred))
```

```
              precision    recall  f1-score   support

           0       1.00      0.56      0.71         9
           1       0.73      1.00      0.85        11

    accuracy                           0.80        20
   macro avg       0.87      0.78      0.78        20
weighted avg       0.85      0.80      0.79        20
```

Рисунок 29 – Метрики реализации TF-IDF на тестовых данных

3. LDA – метод, позволяющая выделить скрытые тематики среди перечня документов и выявить наиболее значимые слова для данной темы. Для данного алгоритма было решено разработать несколько моделей, различающихся в количестве тематик, которые необходимо найти (задается в параметрах модели). Первый вариант ищет две тематики: предполагается, что в одной будут использоваться термины из IT-области, а в другой, соответственно, термины из других областей, однако, результат получился неудачным, так как обе тематики, по итогу, связаны с IT, а многие слова из других областей были просто выброшены (рисунок 30):

```
(0, '0.020*"социология" + 0.018*"анализ" + 0.014*"интерфейса" + 0.013*"пользовательского" + 0.010*"теория" + 0.007*"взаимодействия" + 0.006*"организации" + 0.006*"графического" + 0.005*"фотометрический" + 0.005*"рельефа"')
(1, '0.019*"систем" + 0.015*"технологий" + 0.013*"управления" + 0.013*"информационных" + 0.012*"разработка" + 0.011*"системы" + 0.011*"медицина" + 0.010*"интерфейсов" + 0.009*"информации" + 0.009*"защиты"')
```

Рисунок 30 – Разделение набора данных по двум тематикам

Второй вариант включает в себя разделение уже на шесть классов: подобласти в IT (UI, машинное обучение, безопасность, коммерция и внедрение в производство) и области, не связанные с IT. Результат работы представлен на рисунке 31.

```
(0, '0.064*"анализ" + 0.029*"методы" + 0.015*"рельефа" + 0.015*"восстановления" + 0.015*"лице" + 0.015*"операциях" + 0.015*"поверхности" + 0.015*"лица" + 0.015*"фотометрический" + 0.009*"сравнительный"')
(1, '0.041*"исследование" + 0.028*"российских" + 0.025*"электронных" + 0.024*"примере" + 0.020*"трансакционных" + 0.020*"современные" + 0.020*"снижения" + 0.020*"технологии" + 0.020*"предприятиях" + 0.020*"издержек"')
(2, '0.064*"систем" + 0.045*"информационных" + 0.042*"медицина" + 0.028*"технологий" + 0.022*"россии" + 0.021*"пользовательского" + 0.019*"бизнеса" + 0.019*"использование" + 0.019*"развитии" + 0.016*"классификация"')
(3, '0.048*"разработка" + 0.042*"системы" + 0.035*"теория" + 0.032*"социология" + 0.026*"организации" + 0.023*"рынка" + 0.023*"услуг" + 0.022*"создание" + 0.017*"автоматизированной" + 0.017*"взаимодействия"')
(4, '0.044*"управления" + 0.031*"моделирование" + 0.026*"интерфейсов" + 0.025*"условиях" + 0.023*"информации" + 0.023*"защиты" + 0.018*"использованием" + 0.018*"оценка" + 0.016*"средствами" + 0.014*"технологий"')
(5, '0.034*"совершенствование" + 0.022*"современных" + 0.018*"социальный" + 0.017*"аспекты" + 0.016*"лечения" + 0.016*"нервов" + 0.016*"нижних" + 0.016*"верхних" + 0.016*"конечностей" + 0.016*"травм"')
```

Рисунок 31 – Разделение набора данных по шести тематикам

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Можно сделать вывод, что использование LDA эффективно проявит себя при поиске подобластей уже выбранной темы, так как использование данного алгоритма для выделения целевой темы является не лучшим его применением, что было показано на рисунках 30 и 31. В качестве эксперимента, было решено применить данную модель только к тем работам, которые имеют подобласть в IT. В целом, имеется разделение по нескольким темам, но для лучшего результата требуется увеличение количества анализируемых работ, так как выделенные тематики заметно пересекаются между собой. Результат разбиения по тематикам представлен на рисунке 32.

```
(0, '0.063*"разработка" + 0.058*"технологий" + 0.047*"системы" + 0.027*"рынка" + 0.027*"услуг" + 0.019*"автоматизированной" + 0.017*"региональной" + 0.017*"информационной" + 0.017*"хозяйстве" + 0.017*"платежной"')
(1, '0.077*"систем" + 0.045*"теория" + 0.045*"организации" + 0.040*"исследование" + 0.039*"интерфейса" + 0.039*"пользовательского" + 0.036*"информационных" + 0.024*"информации" + 0.024*"защиты" + 0.015*"эффективность"')
(2, '0.054*"оценка" + 0.049*"качества" + 0.029*"проектирования" + 0.020*"критерии" + 0.018*"оценки" + 0.015*"информационная" + 0.014*"приложений" + 0.012*"удобства" + 0.009*"тайлов" + 0.009*"использования"')
(3, '0.069*"управления" + 0.029*"классификация" + 0.027*"моделирование" + 0.024*"совершенствование" + 0.022*"инвестиций" + 0.022*"эффективность" + 0.022*"информатизацию" + 0.020*"средствами" + 0.016*"графического" + 0.014*"технологии"')
(4, '0.064*"интерфейсов" + 0.040*"электронных" + 0.039*"методов" + 0.028*"пользовательских" + 0.017*"сети" + 0.017*"анализ" + 0.017*"безопасных" + 0.017*"платежей" + 0.017*"интернет" + 0.017*"теоретический"')
```

Рисунок 32 – Разделение подобластей

4. LSI – метод для определения тематик в текстовых данных, однако, в отличие от LDA для нахождения семантических связей он использует матрицу «слово-текст», когда LDA опирается на вероятностную модель, то есть находит распределение вероятностей слов в каждой теме. В целом, разработка модели по данному алгоритму не имеет отличий от LDA. На рисунках 33-35 представлены результаты по тем же вариантам моделей. Увеличение количества элементов в изначальном наборе данных также повысит эффективность модели.

```
(0, '0.503*"интерфейса" + 0.475*"пользовательского" + 0.332*"качества" + 0.259*"интерфейсов" + 0.228*"пользовательских" + 0.209*"оценка" + 0.195*"оценки" + 0.152*"систем" + 0.118*"обеспечения" + 0.109*"информационных"')
(1, '0.525*"интерфейсов" + 0.475*"пользовательских" + -0.323*"интерфейса" + -0.264*"пользовательского" + 0.249*"оценки" + 0.129*"адаптивных" + -0.125*"качества" + 0.113*"интеллекта" + 0.113*"искусственного" + 0.110*"графических"')
```

Рисунок 33 – Разделение набора данных по двум тематикам

```
(0, '0.502*"интерфейса" + 0.448*"пользовательского" + 0.365*"качества" + 0.264*"интерфейсов" + 0.243*"пользовательских" + 0.222*"оценка" + 0.179*"оценки" + 0.166*"систем" + 0.113*"обеспечения" + 0.104*"информационных"')
(1, '0.530*"интерфейсов" + 0.494*"пользовательских" + -0.331*"интерфейса" + -0.251*"пользовательского" + 0.193*"оценки" + -0.168*"качества" + 0.142*"адаптивных" + 0.117*"графических" + 0.111*"интеллекта" + 0.111*"искусственного"')
(2, '0.511*"систем" + 0.324*"информационных" + 0.241*"россии" + -0.228*"качества" + 0.194*"государственных" + 0.181*"мчс" + -0.133*"адаптивных" + 0.130*"оценки" + -0.130*"интерфейсов" + 0.126*"управления"')
(3, '0.526*"оценки" + 0.292*"качества" + -0.195*"системы" + -0.158*"разработка" + -0.147*"систем" + -0.147*"адаптивных" + 0.134*"формирование" + 0.130*"группы" + 0.130*"экспертов" + -0.127*"пользовательского"')
(4, '0.277*"импульсного" + 0.277*"движения" + 0.277*"спиновых" + 0.277*"оболочке" + 0.277*"молекулярные" + 0.277*"модельной" + 0.277*"меток" + 0.277*"мембране" + 0.277*"эпр" + 0.277*"ближайшей"')
(5, '0.277*"мероприятий" + 0.277*"питомниках" + 0.277*"фомоза" + 0.277*"сосны" + 0.277*"посадочного" + 0.277*"обыкновенной" + 0.277*"обоснование" + 0.277*"беларуси" + 0.277*"европейской" + 0.277*"защите"')
```

Рисунок 34 – Разделение набора данных по шести тематикам

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

```
(0, '0.504*"интерфейса" + 0.448*"пользовательского" + 0.358*"качества" + 0.263*"интерфейсов" + 0.242*"пользовательских" + 0.223*"оценка" + 0.183*"оценки" + 0.168*"систем" + 0.115*"обеспечения" + 0.106*"информационных"')
(1, '0.527*"интерфейсов" + 0.494*"пользовательских" + -0.334*"интерфейса" + -0.255*"пользовательского" + 0.194*"оценки" + -0.160*"качества" + 0.141*"адаптивных" + 0.120*"графических" + 0.111*"интеллекта" + 0.111*"искусственного"')
(2, '0.564*"систем" + 0.336*"информационных" + 0.225*"государственных" + -0.216*"качества" + 0.158*"россии" + 0.149*"управления" + 0.145*"мчс" + -0.135*"оценка" + 0.131*"модели" + 0.131*"сервисов"')
(3, '0.593*"оценки" + 0.193*"формирование" + 0.184*"группы" + 0.184*"экспертов" + 0.174*"качества" + -0.168*"адаптивных" + 0.151*"метод" + -0.151*"искусственного" + -0.151*"интеллекта" + -0.145*"системы"')
(4, '-0.473*"качества" + 0.252*"пользовательского" + 0.211*"элементов" + -0.200*"систем" + -0.146*"программного" + 0.132*"сетки" + 0.132*"сложности" + 0.132*"алгоритма" + 0.132*"расположения" + 0.132*"определения"')
```

Рисунок 35 – Разделение подобластей

5. Word2vec – это техника обработки естественного языка, использующая нейронные сети на большом объеме текста, где каждое слово рассматривается как целевая переменная в задаче обучения с учителем, используется для создания векторных представлений слов в высокомерном пространстве, где позиция вектора отражает значение слова. Это достигается за счет обучения нейронной сети. Сеть обучается предсказывать контекстные слова, окружающие каждое целевое слово, и веса сети используются для создания векторов слов. Существует две основные архитектуры для word2vec: модель непрерывного мешка слов (CBOW) и модель skip-gram. В модели CBOW сеть предсказывает целевое слово на основе его контекстных слов, в то время как в модели skip-gram сеть предсказывает контекстные слова, заданные целевым словом. Особенностью word2vec является его возможность обновления векторов при добавлении новых слов (токенов), что делает данный вариант довольно гибким при использовании. На рисунке 36 показан вывод слов, наиболее близких по весу к переданному в качестве параметра. На основе данного результата и особенности word2vec постоянно совершенствоваться за счет пополнения слов эта модель будет считаться одной из приоритетных для использования.

```
[51] print(wordmodel.wv.most_similar(['информационная'], topn=5))

    [('пользовательского', 0.37489834427833557), ('медицина', 0.3196730613708496), ('обыкновенной', 0.3049739897251129), ('систем', 0.2852994501590729), ('ui', 0.28351712226867676)]
```

Рисунок 36 – Вывод наиболее похожих слов

6. LSTM – модификация рекуррентной нейронной сети, особенность которой заключается в возможности обработки больших текстов. Особенностью разработки модели является отсутствие необходимости в предварительной обработке текста, так как LSTM может самостоятельно

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

извлекать признаки из текста и использовать их для классификации или генерировать последовательность на основе данного контекста, однако, токенизация также остается необходимой. На рисунке 37 представлена архитектура сети, состоящей из нескольких слоев. Слой Embedding служит для преобразования целочисленных индексов слов в векторные представления, которые могут быть обработаны нейронной сетью, кроме того, использование данного слоя позволяет уменьшить размерность входного пространства и ускорить обучение модели, так как векторные представления имеют более компактный формат. Слой Bidirectional(LSTM) представляет из себя двунаправленную LSTM, которая обрабатывает контекст в прямом и обратном порядке. Последний слой Dense служит получения выходных данных, в данном случае которые являются значением принадлежности работы к области информационных технологий (1 – да, 0 – нет). На рисунке 38 показаны значения метрик, значения которых говорят о том, что нейросеть имеет высокие показатели определения принадлежности работы к области IT.

Рисунок 37 – Архитектура нейронной сети LSTM

```
lstm = Sequential()
lstm.load_weights(model_lstm_save_path)
lstm.evaluate(x_test, y_test, verbose=1)
```

```
1/1 [==============================] - 1s 989ms/step - loss: 0.2642 - accuracy: 0.9000 - precision: 0.9000 - recall: 0.9000
[0.26419854164123535,
 0.8999999761581421,
 0.8999999761581421,
 0.8999999761581421]
```

Рисунок 38 – Показатели метрик на тестовой выборке

В результате имеются следующие варианты моделей:

- TF-IDF (метрика accuracy – 0.8).
- LSTM (метрика accuracy – 0.9).
- Doc2vec (отсутствуют стандартные метрики качества, является подходящим для работ, терминология которых не встречалась в изначальных текстовых данных).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

В рамках реализации проекта решено использовать модель LSTM, однако, использование doc2vec является более корректным, так как это решает проблему использования в названиях неизвестных терминов. Относительно моделей LSI и LDA можно сказать, что они более применимы для распределения работ по подобластям, однако, также важно количество изначальных данных.

При поступлении новой работы со стороны сервера она проходит обработку модели и получает значение метки предсказания. Работа получает обновление в базе данных, где обновляется данное свойство, которое изначально не было определено. При совпадении с целевой меткой и принадлежности работы к общей тематике она публикуется на информационной платформе.

Для подключения к базе данных используется метод connect из модуля psycopg2, в параметры которого передаются название базы данных, пароль для доступа к ней, имя пользователя-владельца, имя сервера, на котором запущен PostgreSQL, и номер порта (рисунок 39).

```
try:
    db = psycopg2.connect(dbname='scientific_works',
                          user='postgres',
                          password='123456789',
                          host='localhost',
                          port=5433)
except:
    print('Не удалось подключиться к базе данных')
```

Рисунок 39 – Создание соединения с базой данных

Чтобы появилась возможность выполнять SQL-запросы, необходимо использовать метод объекта соединения cursor. Для получения списка работ будет использоваться метод SELECT, а для обновления свойства pred у определенной работы – метод UPDATE.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

3.2 Разработка серверной части информационной платформы

Для разработки сервера используется язык JavaScript и платформа Node.js, которая использует библиотеку Express.

В первую очередь, необходимо инициализировать сервер. Для этого импортируется модуль express и с его помощью происходит первоначальная инициализация сервера (рисунок 40). Метод listen() запускает прослушивание сервера, то есть становится возможным отправка локальных запросов на указанный порт, значение которого считывается из файла .env, содержащий в себе глобальные константы, либо присваивается в параметрах по умолчанию. Доступ к файлу .env осуществляется через модуль dotenv.

```
const express = require("express");
require("dotenv").config();

const app = express();

const start = async () => {
  try {
    app.listen(process.env.PORT || 5000, () => console.log("listen port:", process.env.PORT || 5000));
  } catch (error) {
    console.log(error);
  }
};

start();
```

Рисунок 40 – Инициализация сервера

Для взаимодействия базой данных необходимо осуществить соединение с ней. Сделать это можно при помощи класса Sequelize из одноименного модуля. Чтобы использовать объект соединения в разработке, необходимо предоставить возможность другим файлам сервера использовать его при помощи module.exports, которому присваивается новый экземпляр класса Sequelize, в конструктор которого передается имя базы данных, имя пользователя-владельца, пароль, диалект, то есть указывание используемой базы данных, сервер и порт, где она запущена. Пример создания такого экземпляра представлен на рисунке 41. Для подключения сервера внутри инициализированного приложения используются методы объекта соединения

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

authenticate(), который производит подключение при помощи данных, переданных во время создания экземпляра, и sync(), который реагирует на изменения в базе данных.

```
module.exports = new Sequelize(
  process.env.DB_NAME,
  process.env.DB_USER,
  process.env.DB_PASSWORD,
  {
    dialect: "postgres",
    host: process.env.DB_HOST,
    port: process.env.DB_PORT,
  },
);
```

Рисунок 41 – Экземпляр объекта соединения

При помощи объекта соединения создаются таблицы works и users, которые хранят информацию о работах и пользователях соответственно. Во время создания таблицы необходимо определить ее имя, поля, их тип и дополнительные свойства. Для таблицы works были определены следующие поля:

- id – уникальный идентификатор, является первичным ключом.
- name – название работы.
- annotation – описание/аннотация к работе.
- subtype – подтип работы, если имеется.
- is_IT_area – принадлежность к IT-области.
- pred – принадлежность к IT-области, основанное на предсказании модели машинного обучения.

Таблица users хранит в себе основную информацию о пользователе: электронную почту, пароль, роль пользователя и его id.

Также между таблицами users и works установлена связь «один-ко-многим» (у одного пользователя может быть создано несколько работ).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

Подробное описание таблиц изображено на рисунке 42.

```
const sequelize = require("../db");
const {DataTypes} = require("sequelize");

const User = sequelize.define("user", {
  id: {type: DataTypes.INTEGER, primaryKey: true, autoIncrement: true},
  email: {type: DataTypes.STRING, unique: true},
  password: {type: DataTypes.STRING},
  role: {type: DataTypes.STRING, defaultValue: "USER"}
});

const Work = sequelize.define("work", {
  id: { type: DataTypes.INTEGER, primaryKey: true, autoIncrement: true },
  name: { type: DataTypes.STRING, unique: true },
  annotation: { type: DataTypes.STRING, unique: true },
  subtype: { type: DataTypes.STRING, defaultValue: "" },
  is_IT_area: { type: DataTypes.BOOLEAN, defaultValue: Boolean(1) },
  pred: { type: DataTypes.BOOLEAN, defaultValue: Boolean(0) },
});

User.hasMany(Work);
Work.belongsTo(User);

module.exports = {
  User,
  Work,
};
```

Рисунок 42 – Описание таблиц в базе данных

Для осуществления взаимодействия с элементами в базе данных на серверной стороне необходимо описать методы HTTP-запросов и маршруты – адреса сервера, по которым будет обращаться клиентская часть для вызова этих методов. Алгоритм создания подобных маршрутов следующий:

1. Создается файл маршрутов для каждой таблицы, в котором описывается используемый HTTP-метод, адрес, по которому происходит обращение, и функция, изменяющая базу данных. Для удобства описания функций создается класс, который определяет методы для соответствующей таблицы (рисунок 43).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

2. Создается файл всех маршрутов, который объединяет их в одном объекте (рисунок 44). Сделано это для удобства описания методов, так как отсутствует лишний шаблонный, то есть повторяющийся, код.

3. В коде инициализации приложения определяется коренной маршрут, который является началом для всех остальных.

```
class WorkController {
  async create(req, res) {
    const data = req.body;
    const work = await Work.create({ name: data.name });
    return res.json(work);
  }

  async getOne(req, res) {
    const { name, id } = req.params;
    const work = await Work.findOne({ where: { id } });
    return res.json(work);
  }

  async getAll(req, res) {
    const works = await Work.findAll();
    return res.json(works);
  }

  async delete(req, res, next) {
    const { id } = req.params;
    const work = await Work.findOne({ where: { id } });
    if (!work) {
      return next(ApiError.badRequest("Несуществующий id"));
    }
    await Work.destroy({ where: { id } });
    return res.json({ message: `${work.name} was deleted` });
  }
}
```

Рисунок 43 – Методы, взаимодействующие с таблицей works

Для таблиц works и users были созданы классы WorkController и UserController. WorkController реализует методы создания и удаления работы, получения одной или их списка. В классе UserController разработаны методы авторизации, регистрации и метод, который проверяет, авторизован ли

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

пользователь, данные методы используют технологию JSON Web Token [], суть которой заключается в следующем:

1. Происходит регистрация или авторизация пользователя.
2. Данные, введенные пользователем, кодируются в виде специальной строки. В рамках проекта данная строка хранит информацию об электронной почте, роли и уникальном идентификаторе пользователя. С помощью данной строки происходит проверка того, какой пользователь на данный момент пользуется платформой. Общий вид JSON Web Token представлен на рисунке 45, данные, необходимые для проверки авторизации, находятся в секции payload.

```
const router = new Router();

router.post("/", workController.create);
router.get("/", workController.getAll);
router.get("/:id", workController.getOne);
router.delete("/:id", checkRole("ADMIN"), workController.delete);

module.exports = router;
```

Рисунок 44 – HTTP-методы для works

Также на рисунке 44 у метода delete имеется дополнительная функция checkRole(), проверяющая роль зарегистрированного пользователя. Такая функция называется middleware – функция, служащая предварительным обработчиком, которая позволяет ввести некоторые ограничения или, наоборот, дополнительные возможности. В рамках реализации серверной составляющей проекта были разработаны следующие middleware:

- errorHandler() – middleware, позволяющий использовать вывод собственных ошибок во время выполнения запросов, при этом если ошибка является не из числа обрабатываемых, то об этом также выведется сообщение.
- authMiddleware() – middleware, проверяющий корректность JSON Web Token авторизированного пользователя.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- checkRole() – middleware, который проверяет роль авторизованного пользователя.



Рисунок 45 – Структура JSON Web Token

На рисунке 46 показана итоговая структура приложения, в которой следует выделить следующие особенности:

- Метод cors() позволяет осуществлять запросы к серверной составляющей со стороны клиента, изначально действует политика Cross-Origin Resource Sharing, которая блокирует запросы со стороннего ресурса (сервер и клиент имеют разные порты: 5400 и 3000).
- Метод json() из библиотеки Express позволяет производить обработку ответов на запросы в формате JSON [].
- Важным является порядок объявления middleware errorHandler, так как если добавить после него какой-либо функционал через метод use(), то он не сможет обработать добавленную область, так как она была вызвана позже.

Запуск сервера происходит при помощи команды nodemon <имя файла, в котором определен сервер>. nodemon – модуль, который позволяет отслеживать изменения в коде в режиме реального времени и обновлять запущенный сервер, чего не может позволить себе команда node: при любых изменениях необходим перезапуск сервера.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

```
const app = express();
app.use(cors());
app.use(express.json());
app.use("/api", router);
app.use(errorHandler);

const start = async () => {
  try {
    await sequelize.authenticate();
    await sequelize.sync();
    app.listen(process.env.PORT || 5000, () => console.log("listen port:", process.env.PORT || 5000));
  } catch (error) {
    console.log(error);
  }
};

start();
```

Рисунок 46 – Итоговая структура приложения сервера

3.3 Разработка клиентской части информационной платформы

Инициализация клиентской части была выполнена при помощи команды create-react-app, создающей каталог с базовым функционалом React-приложения.

Важной составляющей является определение навигации внутри проекта между компонентами. Для этого главный компонент (App) возвращает компонент BrowserRouter [54] из библиотеки react-router-dom, внутри которого настроена навигация при помощи таких составляющих, как Routes (создает группу маршрутов) и Route (определение маршрута), в который передается страница-компонент и путь до нее, который будет прописан в адресной строке. На рисунке 47 представлена настроенная маршрутизация информационной платформы. Некоторые Route объявлены через фигурные скобки: это нужно для того, чтобы задать условие отображения элемента (одна из особенностей JSX-синтаксиса). В данном случае проверяется, авторизован пользователь или нет, информацию о котором можно получить при помощи использования специального метода useContext, позволяющий использовать значения глобально, по всему проекту (это избавляет от известной проблемы prop drilling, когда необходимо использовать компонент-посредник для передачи значения от родительского элемента до более глубокого уровня). Сам Context

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

определяется внутри главного компонента App. Вместе с useContext используется стейт-менеджер MobX [55], позволяющий настроить взаимодействие с глобальным контекстом, а также улучшающий производительность приложения, так как регулирует обновление (ререндеринг) компонентов при помощи метода observer().

```
export const AppRouter = observer(() => {
  const {user} = useContext(Context);
  console.log(user);
  return (
    <>
      <Routes>
        {!user.isAuth && <Route path='/login' element={<Auth />} />}
        {user.isAuth && <Route path='/' element={<WorksPage />} />}
        <Route element={<Auth />} path='registration' />
        <Route path='/' element={<Auth />} />
        <Route path='/works' element={<WorksPage />} />
        <Route path='/works/:id' element={<WorkInfoPage />} />
        <Route
          path='*'
          element={
            <h1 style={{ fontSize: "32px", margin: "7px" }}>
              404: Page does not exist
            </h1>
          }
        />
      </Routes>
    </>
```

Рисунок 47 – Настроенная маршрутизация

Также MobX выступает в роли хранилища данных глобального объекта состояния. Для этого создается файл под каждое глобальный объект, который реализует соответствующий ему класс, предоставляющий значения свойств экземпляра при его инициализации, методы изменения и получения свойств объекта. На рисунке 48 изображен класс состояния, в котором определяется

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

текущий пользователь, а также свойство isAuth, которое позволяет получить доступ к некоторым функциям.

```
export default class UserStore {
  private _isAuth: boolean;
  private _user: IUser;
  constructor() {
    this._isAuth = false;
    this._user = {} as IUser;
    makeAutoObservable(this);
  }

  setIsAuth(bool: boolean) {
    this._isAuth = bool;
  }

  setUser(user: IUser) {
    this._user = user;
  }

  get isAuth() {
    return this._isAuth;
  }

  get user() {
    return this._user;
  }
}
```

Рисунок 48 – Класс состояния

При помощи глобальных состояний определяется список источников. В компоненте WorksPage используется запрос на получение работ сервера, а при помощи метода useEffect получение данных происходит корректно и не конфликтует с другими процессами. Формирование списка подобластей формируется в зависимости от того, какие работы были получены со стороны сервера, за счет метода useState происходит обновление списка работ при изменении значения списка подобластей.

Информационная платформа имеет ролевую модель, то есть несколько категорий пользователей, которые потенциально будут использовать платформу:

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

- Незарегистрированный пользователь – имеет функционал просмотра и сортировки работ по подобласти (рисунок 49).

- Зарегистрированный пользователь – имеет тот же функционал, а также возможность добавить свою научную работу (рисунок 50).

- Модератор – кроме добавления работ имеет возможность удалять их, если они прошли фильтрацию модели машинного обучения, но при ручной проверке оказались неподходящими по основной тематике (рисунок 51).

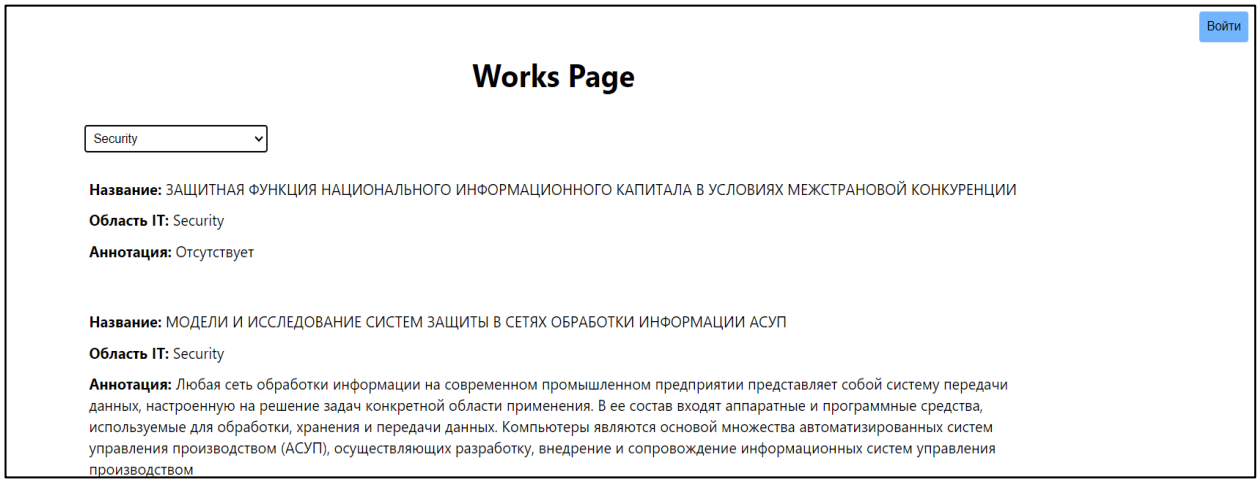


Рисунок 49 –Платформа для незарегистрированного пользователя

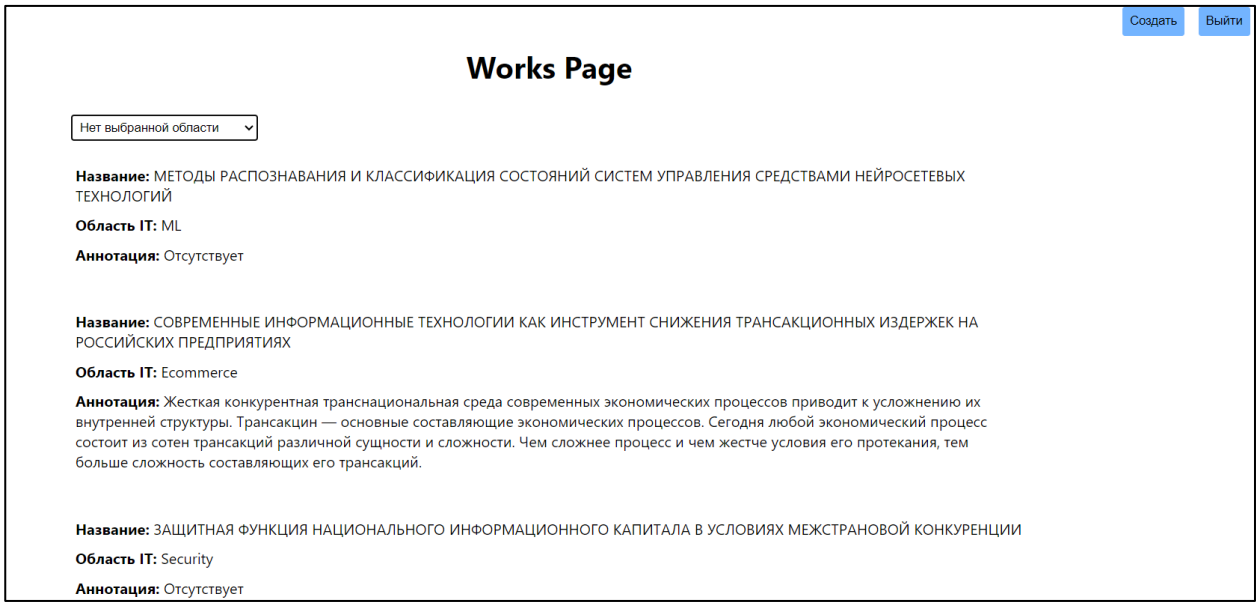


Рисунок 50 –Платформа для зарегистрированного пользователя

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

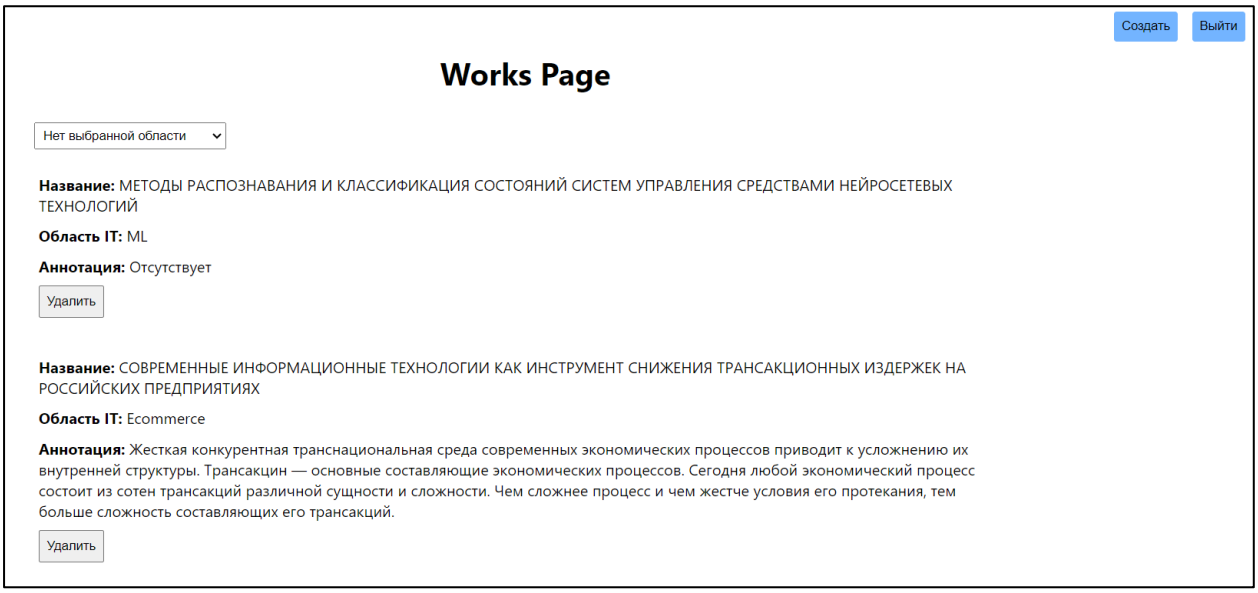


Рисунок 51 – Платформа для модератора

При создании работы она добавляется в базу данных, из которой извлекается при помощи библиотеки Psycopg2 и отправляется на проверку через обученную модель. В результате обучения ей присваивается значение свойства is_IT_area: если значение при приведении типа к логическому является true, то оно будет отображаться на информационной платформе. На рисунке 28 представлена архитектура проекта, желтая стрелка демонстрирует добавление работы.

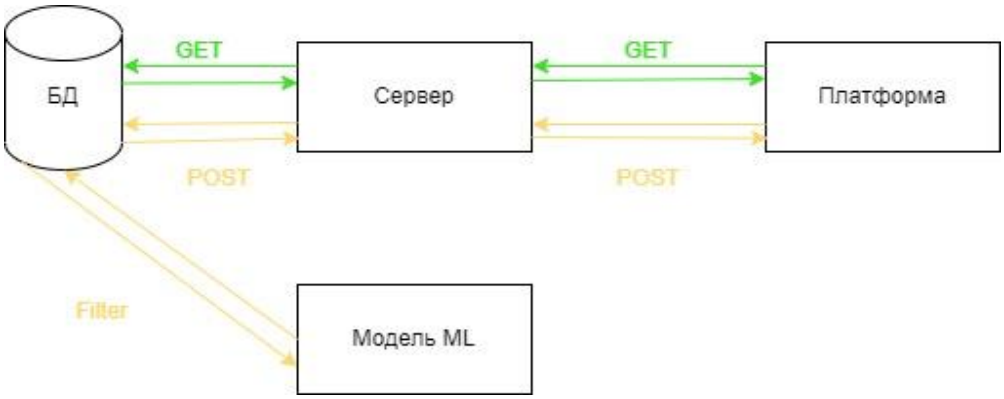


Рисунок 52 – Взаимодействие между компонентами проекта

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

3.4 Результаты и выводы к главе

Данная глава является заключительной и ее результатом служит разработанная информационная платформа, которая в качестве предварительной фильтрации данных использует модель машинного обучения.

Были разработаны модели машинного обучения, а также проведено их сравнение между собой с использованием метрик. В результате данного сравнения было получено следующее:

- По показателям метрик лучшей является двунаправленная модель LSTM.
- Для обработки большего количества данных, а также учета новых терминов подойдет модель Word2vec, так как, в отличие, например, от LSTM данная модель может обновлять свой словарь при передаче ей новых слов.
- LSI и LDA подходят для выявления подобластей основной темы, однако, также нуждаются в большем изначальном объеме данных.

Был разработан сервер, являющийся «посредником» между клиентской частью информационной платформы и базой данных, запросы к которой она и осуществляет через сервер.

Была разработана клиентская составляющая в виде сайта с ролевой моделью, обеспечивающей различный функционал для групп пользователей (незарегистрированные, зарегистрированные и модераторы). Сам сайт служит для отображения работ, прошедших фильтрацию моделью машинного обучения.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

ЗАКЛЮЧЕНИЕ

В данной магистерской диссертации была проведена работа по изучению и применению методов машинного обучения для фильтрации текстовых данных. Одним из ключевых вызовов, которые стоят перед исследователями в данной области, является необходимость автоматического анализа большого объема текстовых данных, что может быть достигнуто при помощи использования методов машинного обучения.

В первой главе работы был проведен анализ научных исследований, связанных с использованием методов машинного обучения для фильтрации данных. Было выявлено, что данный подход является эффективным и может быть использован для автоматизации процесса фильтрации текстовых данных. Кроме обзора научных статей был предоставлен перечень терминов и определений, которые были использованы в процессе выполнения работы.

Во второй главе был проведен обзор инструментария, который представляет собой необходимый набор средств для разработки модели машинного обучения для фильтрации текстовых данных и информационной платформы, предназначенной для отображения полученных результатов. Были рассмотрены различные инструменты: tensorflow.keras, genism, sklearn, которые могут быть использованы для разработки модели машинного обучения, и ReactJS, TypeScript, Node.js, применяемые для фронтенд- и бэкенд-разработки.

В третьей главе были разработаны модели машинного обучения, фильтрующие текстовые данные по целевой метке и информационная платформа. Среди моделей была выбрана лучшая при помощи метрик. Информационная платформа была разработана для отображения результатов работы модели и была успешно интегрирована с моделью машинного обучения через базу данных.

Таким образом, полученная модель машинного обучения и информационная платформа могут быть использованы для автоматизации

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

процесса фильтрации текстовых данных в различных областях, например, в тематических электронных библиотеках. Эти результаты открывают новые возможности для улучшения качества фильтрации текстовых данных и повышения эффективности работы с большим объемом информации.

Также стоит отметить, что данная работа является актуальной и перспективной для дальнейших исследований в области машинного обучения. В дальнейшем можно рассмотреть возможности для оптимизации модели машинного обучения, улучшения ее точности и скорости работы, а также расширения функционала информационной платформы.

В целом, результаты данной магистерской диссертации позволяют сделать вывод о том, что использование методов машинного обучения для фильтрации текстовых данных является эффективным подходом, который может быть применен для автоматизации процесса фильтрации большого объема информации.

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. SciVal [Electronic resource]. – 2023. – URL: https://www.scival.com/ (дата обращения: 07.03.2023).

2. Академия Google [Электронный ресурс]. – 2023. – URL: https://scholar.google.com/ (дата обращения: 11.04.2023).

3. Все статьи подряд / Хабр [Электронный ресурс]. – 2023. – URL: https://habr.com/ru/all/ (дата обращения: 05.04.2023).

4. Инновация – Википедия [Электронный ресурс]. – 2023. – URL: https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D0%BD%D0%BE%D0%B2%D0%B0%D1%86%D0%B8%D1%8F (дата обращения: 16.04.2023).

5. Jira | Программное обеспечение для отслеживания задач и проектов [Электронный ресурс]. – 2023. – URL: https://www.atlassian.com/ru/software/jira (дата обращения: 05.05.2023).

6. Трансфер технологий – Википедия [Электронный ресурс]. – 2023. – URL: https://ru.wikipedia.org/wiki/%D0%A2%D1%80%D0%B0%D0%BD%D1%81%D1%84%D0%B5%D1%80_%D1%82%D0%B5%D1%85%D0%BD%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D0%B9 (дата обращения: 16.04.2023).

7. 6.2 Цифровые проекты и платформы [Электронный ресурс]. – 2022. – URL: https://strategy.cdto.ranepa.ru/6-2-cifrovye-proekty-i-platformy#:~:text=%D0%98%D1%82%D0%B0%D0%BA%2C%20%D1%81%20%D1%82%D0%BE%D1%87%D0%BA%D0%B8%20%D0%B7%D1%80%D0%B5%D0%BD%D0%B8%D1%8F%20%D0%98%D0%A2,%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B5%20%D0%B7%D0%B0%D0%B4%D0%B0%D1%87%D0%B8%20%D0%B2%20%D0%B0%D0%B2%D1%82%D0%BE%D0%BC%D0%B0%D1%82%D0%B8%D0%B7%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%BD%D0%BE%D0%BC%20%D1%80%D0%B5%D0%B6%D0%B8%D0%BC%D0%B5 (дата обращения: 17.04.2023).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

8. SciVal Руководство по регистрации и работе в системе. – URL: https://etu.ru/assets/files/Faculty-FEM/ban/scival_rukovodstvo-po-rabote.pdf (дата обращения: 07.03.2023).

9. eLIBRARY.RU - НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА [Электронный ресурс]. – 2023. – URL: https://www.elibrary.ru/defaultx.asp (дата обращения: 20.04.2023).

10. Ding L. Multimodal transport information sharing platform with mixed time window constraints based on big data //Journal of Cloud Computing. – 2020. – Т. 9. – С. 1-11.

11. Tinyakova V. I., Morozova N. I., Karyagina T. V. Sustainable Development of the Digital Economy: Global Challenges to Russia's National Security and Prospects for Its Innovative Breakthrough //Digital Technologies and Institutions for Sustainable Development. – Cham : Springer International Publishing, 2022. – С. 471-475.

12. Yang S. et al. Semantic inference on clinical documents: combining machine learning algorithms with an inference engine for effective clinical diagnosis and treatment //IEEE Access. – 2017. – Т. 5. – С. 3529-3546.

13. Gupta N., Gupta B. Machine learning approach of semantic mapping in polystore health information systems //Int J Comput Inf Syst Ind Manage Appl. – 2021. – Т. 13. – С. 222-232.

14. Hsu I. C., Chang C. C. Integrating machine learning and open data into social Chatbot for filtering information rumor //Journal of Ambient Intelligence and Humanized Computing. – 2021. – Т. 12. – С. 1023-1037.

15. Dafaallah D. E., Hashim A. S. Sentiment analysis techniques to analyze hse situational awareness at oil and gas platforms using machine learning //Indian Journal of Computer Science and Engineering. – 2020. – Т. 11. – №. 5. – С. 640-645.

16. ФИПС - Федеральное государственное бюджетное учреждение Федеральный институт промышленной собственности [Электронный ресурс]. – 2023. – URL: https://www.fips.ru/ (дата обращения: 03.05.2023).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

17. eLIBRARY.RU - Поиск патентов [Электронный ресурс]. – 2023. – URL: https://www.elibrary.ru/patents.asp (дата обращения: 03.05.2023).

18. Яндекс.Патенты – поиск по патентным документам [Электронный ресурс]. – 2023. – URL: https://yandex.ru/patents (дата обращения: 05.05.2023).

19. Google Patents [Electronic resource] – 2023. – URL: https://patents.google.com/ (дата обращения: 06.05.2023).

20. Информационно-поисковая система [Электронный ресурс]. – 2023. – URL: https://www.fips.ru/iiss/document.xhtml?faces-redirect=true&id=d1a728a4b497590a04bc60888c277c6c (дата обращения: 03.05.2023).

21. СВЯЗАННАЯ ЭКСПЕРТНАЯ СИСТЕМА [Электронный ресурс]. – 2023. – URL: https://www.elibrary.ru/item.asp?id=38457559 (дата обращения: 04.05.2023).

22. RU2743932C2 - СПОСОБ И СЕРВЕР ДЛЯ ПОВТОРНОГО ОБУЧЕНИЯ АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ - Яндекс.Патенты [Электронный ресурс]. – 2023. – URL: https://yandex.ru/patents/doc/RU2743932C2_20210301 (дата обращения: 04.05.2023).

23. 9 ключевых алгоритмов машинного обучения простым языком / Хабр [Электронный ресурс] – 2023. – URL: https://habr.com/ru/articles/509472/ (дата обращения: 07.05.2023).

24. What is the main difference between RNN and LSTM | NLP | RNN vs LSTM – Data Science Duniya [Electronic resource]. – 2023. – URL: https://ashutoshtripathi.com/2021/07/02/what-is-the-main-difference-between-rnn-and-lstm-nlp-rnn-vs-lstm/ (дата обращения: 11.05.2023).

25. scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation [Electronic resource]. – 2023. – URL: https://scikit-learn.org/stable/ (дата обращения: 08.05.2023).

26. Gensim: Topic modelling for humans [Electronic resource]. – 2022. – URL: https://radimrehurek.com/gensim/ (дата обращения: 16.05.2023)

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

27. Module: tf.keras | TensorFlow v2.12.0 [Electronic resource]. – 2023. – URL: https://www.tensorflow.org/api_docs/python/tf/keras (дата обращения: 08.05.2023).

28. REST – Википедия [Электронный ресурс]. – 2023. – URL: https://ru.wikipedia.org/wiki/REST (дата обращения: 20.04.2023).

29. HTTP – Википедия [Электронный ресурс]. – 2023. – URL: https://ru.wikipedia.org/wiki/HTTP (дата обращения: 20.04.2023).

30. HTML – Википедия [Электронный ресурс]. – 2023. – URL: https://ru.wikipedia.org/wiki/HTML (дата обращения: 20.04.2023).

31. JavaScript – Википедия [Электронный ресурс]. – 2023. – URL: https://ru.wikipedia.org/wiki/JavaScript (дата обращения: 20.04.2023).

32. ECMA-262 - Ecma International [Electronic resource]. – 2023. – URL: https://www.ecma-international.org/publications-and-standards/standards/ecma-262/ (дата обращения 21.04.2023).

33. Руководство по DOM - Интерфейсы веб API | MDN [Электронный ресурс]. – 2023. – URL: https://developer.mozilla.org/ru/docs/Web/API/Document_Object_Model (дата обращения 21.04.2023).

34. Window - Интерфейсы веб API | MDN [Электронный ресурс]. – 2023. – URL: https://developer.mozilla.org/ru/docs/Web/API/Window (дата обращения 21.04.2023).

35. Введение - Интерфейсы веб API | MDN [Электронный ресурс]. – 2023. – URL: https://developer.mozilla.org/ru/docs/Web/API/Document_Object_Model/Introduction (дата обращения 21.04.2023).

36. Введение в архитектуру клиент-серверных хранилищ [Электронный ресурс]. – 2015. – URL: http://xn--80aaghdqfmgbznzk1h1c8b.xn--p1ai/lec/mdk0202/0/index.html (дата обращения: 13.05.2023).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

37. PostgreSQL: The world's most advanced open source database [Electronic resource]. – 2023. – URL: https://www.postgresql.org/ (дата обращения: 14.05.2023).

38. pgAdmin 4 — pgAdmin 4 7.3 documentation [Electronic resource]. – 2023. – URL: https://www.pgadmin.org/docs/pgadmin4/latest/index.html (дата обращения: 14.05.2023).

39. React [Electronic resource]. – 2023. – URL: https://react.dev/ (дата обращения: 14.05.2023).

40. TypeScript: JavaScript With Syntax For Types. [Electronic resource]. – 2023. – URL: https://www.typescriptlang.org/ (дата обращения: 14.05.2023).

41. Node.js [Electronic resource]. – 2023. – URL: https://nodejs.org/en (дата обращения: 13.05.2023).

42. Express - Node.js web application framework [Electronic resource]. – 2023. – URL: https://expressjs.com/ (дата обращения: 13.05.2023).

43. Psycopg – PostgreSQL database adapter for Python — Psycopg 2.9.6 documentation [Electronic resource]. – 2023. – URL: https://www.psycopg.org/docs/ (дата обращения: 20.05.2023).

44. pandas - Python Data Analysis Library [Electronic resource]. – 2023. – URL: https://pandas.pydata.org/ (дата обращения: 08.05.2023).

45. NLTK :: Natural Language Toolkit [Electronic resource]. – 2023. – URL: https://www.nltk.org/ (дата обращения: 09.05.2023).

46. pandas.DataFrame — pandas 2.0.2 documentation [Electronic resource]. – 2023. – URL: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html (дата обращения: 08.05.2023).

47. NLTK :: nltk.stem.snowball module [Electronic resource]. – 2023. – URL: https://www.nltk.org/api/nltk.stem.snowball.html?highlight=snowball+stemmer#nltk.stem.snowball.SnowballStemmer (дата обращения: 09.05.2023).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

48. NLTK :: nltk.stem.porter module [Electronic resource]. – 2023. – URL: https://www.nltk.org/api/nltk.stem.porter.html?highlight=porterstemmer#nltk.stem.porter.PorterStemmer (дата обращения: 09.05.2023).

49. NLTK :: nltk.stem.wordnet module [Electronic resource]. – 2023. – URL: https://www.nltk.org/api/nltk.stem.wordnet.html?highlight=wordnetlemmatizer#nltk.stem.wordnet.WordNetLemmatizer (дата обращения: 09.05.2023).

50. Lemmatizer · spaCy API Documentation [Electronic resource]. – 2023. – URL: https://spacy.io/api/lemmatizer (дата обращения: 11.05.2023).

51. TypeScript: Documentation - Everyday Types [Electronic resource]. – 2023. – URL: https://www.typescriptlang.org/docs/handbook/2/everyday-types.html#union-types (дата обращения: 15.05.2023).

52. Postman API Platform [Electronic resource]. – 2023. – URL: https://www.postman.com/ (дата обращения: 14.05.2023).

53. Sequelize | Feature-rich ORM for modern TypeScript & JavaScript [Electronic resource]. – 2023. – URL: https://sequelize.org/ (дата обращения: 14.05.2023).

54. BrowserRouter v6.13.0 | React Router [Electronic resource]. – 2023. – URL: https://reactrouter.com/en/main/router-components/browser-router (дата обращения: 15.05.2023).

55. GitHub - mobxjs/mobx: Simple, scalable state management. [Electronic resource]. – 2023. – URL: https://github.com/mobxjs/mobx (дата обращения: 15.05.2023).

56. Module: tf.keras.layers | TensorFlow v2.12.0 [Electronic resource]. – 2023. – URL: https://www.tensorflow.org/api_docs/python/tf/keras/layers (дата обращения: 17.05.2023).

57. Module: tf.keras.losses | TensorFlow v2.12.0 [Electronic resource]. – 2023. – URL: https://www.tensorflow.org/api_docs/python/tf/keras/losses (дата обращения: 17.05.2023).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты

Электронная библиотека по праву и экономике
https://право.информ2000.рф/uchebniki.htm

58. Module: tf.keras.metrics | TensorFlow v2.12.0 [Electronic resource]. – 2023. – URL: https://www.tensorflow.org/api_docs/python/tf/keras/metrics (дата обращения: 17.05.2023).

59. models.lsimodel – Latent Semantic Indexing — genism [Electronic resource]. – 2022. – URL: https://radimrehurek.com/gensim/models/lsimodel.html (дата обращения: 11.05.2023).

60. models.word2vec – Word2vec embeddings — genism [Electronic resource]. – 2022. – URL: https://radimrehurek.com/gensim/models/word2vec.html (дата обращения: 11.05.2023).

61. models.ldamodel – Latent Dirichlet Allocation — genism [Electronic resource]. – 2022. – URL: https://radimrehurek.com/gensim/models/ldamodel.html (дата обращения: 11.05.2023).

62. useContext – React [Electronic resource]. – 2023. – URL: https://react.dev/reference/react/useContext (дата обращения: 13.05.2023).

Вернуться в каталог дипломов по менеджменту
https://право.информ2000.рф/management3/management3.htm
Полные тексты