РЕФЕРАТ

Изм.

Лист

№ докум.

Подпись

Дата

Лист

*\_*

Разраб.

Провер.

Реценз.

Н. Контр.

Утверд.

Лит.

Листов

Пояснительная записка \_\_ стр., \_\_ рис., \_\_ табл.

КЛЮЧЕВЫЕ СЛОВА.

Цель работы – разработка программы автоматического сопоставления названий товаров для продуктивного использования продающими компаниями. Объектом исследования являются исторические данные о продажах компании. Рассматриваются основы моделей обработки естественного языка, приводятся основные источники шума в экспериментальном наборе данных. Рассмотрена гибридная схема с использованием нескольких моделей машинного обучения на основе архитектуры трансформеров. Приведены алгоритмы работы моделей машинного обучения. Оценена точность приведенных алгоритмов.

СОДЕРЖАНИЕ

[РЕФЕРАТ 2](#_Toc129690728)

[СОДЕРЖАНИЕ 3](#_Toc129690729)

[ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ 4](#_Toc129690730)

[ВВЕДЕНИЕ 5](#_Toc129690731)

[ГЛАВА 1. ПРИНЦИП РАБОТЫ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ АРХИТЕКТУРЫ ТРАНСФОРМЕРОВ 6](#_Toc129690732)

[7.11. ВЫВОДЫ 7](#_Toc129690733)

[ЗАКЛЮЧЕНИЕ 8](#_Toc129690734)

[СПИСОК ЛИТЕРАТУРЫ 9](#_Toc129690735)

[ПРИЛОЖЕНИЕ 1 10](#_Toc129690736)

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

*ML* - Machine Learning (машинное обучение)

*AI* - Artificial Intelligence (искусственный интеллект)

*ANN* - Artificial Neural Network (искусственная нейронная сеть)

*RF* - Random Forest (случайный лес)

*KNN* - K-Nearest Neighbors (метод ближайших соседей)

*NB* - Naive Bayes (наивный байесовский классификатор)

*DL* - Deep Learning (глубокое обучение)

*CNN* - Convolutional Neural Network (сверточная нейронная сеть)

*RNN* - Recurrent Neural Network (рекуррентная нейронная сеть)

*NLP* - Natural Language Processing (обработка естественного языка)

*NLTK* - Natural Language Toolkit (инструментарий для работы с естественным языком)

*POS* - Part of Speech (часть речи)

*NER* - Named Entity Recognition (распознавание именованных сущностей)

*LSA* - Latent Semantic Analysis (латентный семантический анализ)

*LDA* - Latent Dirichlet Allocation (латентное размещение Дирихле)

*SVM* - Support Vector Machine (метод опорных векторов)

*ANN* - Artificial Neural Network (искусственная нейронная сеть)

*MLP* - Multi-Layer Perceptron (многослойный персептрон)

*LSTM* - Long Short-Term Memory (долгая краткосрочная память)

ВВЕДЕНИЕ

Для современного бизнеса в России ежегодно проводятся реформы, направленные на увеличение эффективности компаний и роста их количества на территории государства. В условиях стремительно развивающегося бизнеса важно своевременно найти слабое место на пути к развитию и вовремя его устранить. И, если на уровне государства, можно быстро понять, что тормозит бизнес-процессы в целом, то на уровне отдельно взятой корпорации, при рассмотрении бизнес-процессов, все может быть неоднозначно. Множество факторов могут быть годами скрыты от глаз и из эффект на работы компании не явен. В частности, большое влияние оказывает осведомленность сотрудников компании о трудовых обязанностях друг друга, поскольку понять, что процесс выстроен неэффективно можно лишь зная чем занимается каждый участник процесса.

Поскольку бизнес строится людьми и в бизнес-процессах участвуют люди, на определенном уровне абстракции, можно считать, что развитие бизнеса ограничено именно человеческими способностями, например способностью долго сохранять концентрацию или эффективно вести коммуникацию в рамках рабочих процессов.

Все это подталкивает к тому, что необходимы способы увеличения человеческой продуктивности в задачах, где человека не может полностью заменить компьютер. Машинное обучение в задачах подобного рода наиболее органично встраивается в бизнес-процесс.

Одной из проблем, с которой сталкиваются многие компании, является задача ручного сопоставления удобочитаемых названий продуктов с кодами из системы товарооборота (1С, SAP). Это необходимо, поскольку разные дилеры могут продавать один и тот же товар, но описывать его по-разному, что приводит к несоответствиям в данных. Этот процесс может занимать много времени, быть подверженным ошибкам и ресурсоемким, что делает его первым кандидатом на автоматизацию.

Методы обработки естественного языка (NLP) могут применяться для автоматизации задачи сопоставления описаний продуктов со штрих-кодами или любыми другими внутренними номерами, которые представлены в плохо читаемом формате. Алгоритмы NLP можно обучать на наборе данных описаний продуктов и соответствующих им штрих-кодов, чтобы изучать закономерности и взаимосвязи между ними. После обучения эти алгоритмы можно использовать для прогнозирования штрих-кода для данного описания продукта.

Для этой задачи можно использовать несколько подходов, включая классификацию текста, распознавание именованных объектов и извлечение информации. Эти методы можно комбинировать с алгоритмами машинного обучения, такими как метод опорных векторов, деревья решений или нейронные сети, для создания надежной и точной системы сопоставления штрих-кодов.

Автоматизация процесса сопоставления описаний продуктов со штрих-кодами может значительно повысить эффективность, уменьшить количество ошибок и сэкономить ресурсы компаний. Она также может обеспечить более последовательную и надежную систему для отслеживания запасов и управления данными о продуктах. При наличии правильных инструментов и методов NLP может сыграть решающую роль в упрощении этой задачи и улучшении бизнес-операций.

* **Цели и задачи работы**

Цель этого проекта — создать модель машинного обучения для автоматического сопоставления описаний продуктов с внутренними кодами компании. Это поможет упростить процесс управления данными о продуктах и ​​запасами для компаний.

Конкретные цели проекта включают:

1. Подготовка набора данных описаний продуктов и соответствующих им кодов;
2. Предварительная обработка данных, очистка и подготовка к обучению;
3. Изучение и анализ данных, просмотр статистик;
4. Выбор подходящих методов NLP и машинного обучения для решения задачи;
5. Проектирование и обучение модели машинного обучения на подготовленном наборе данных;
6. Оценка производительности модели и ее модернизация по необходимости;
7. Тестирование модели на новых данных для обеспечения ее точности и эффективности.

В процессе достижения этих целей, можно создать надежную и эффективную систему сопоставления описаний продуктов с их кодами, которая поможет компаниям сэкономить время и ресурсы, повысив точность и согласованность данных о продуктах. Кроме того, этот проект может служить основой для дальнейших исследований и разработок в области NLP и машинного обучения, применяемых для управления запасами и обработки данных о продуктах.

ЛИТОБЗОР

* **Краткое введение в NLP и машинное обучение**

Обработка естественного языка (*NLP*) — это область искусственного интеллекта, которая занимается взаимодействием между компьютерами и человеческим языком. Он включает в себя разработку алгоритмов и моделей, которые могут обрабатывать и понимать текстовые или речевые данные на естественном языке.

Машинное обучение — это область искусственного интеллекта, которая включает в себя использование статистических методов и алгоритмов, позволяющих компьютерам учиться на основе данных без явного программирования. Алгоритмы машинного обучения можно обучать на больших наборах данных, чтобы распознавать закономерности и отношения в данных, что позволяет им делать прогнозы или классифицировать новые данные.

*NLP* и машинное обучение тесно переплетены, и алгоритмы машинного обучения часто используются в задачах *NLP* для анализа, классификации и генерации текста на естественном языке. Например, алгоритмы машинного обучения можно использовать для анализа тональности, классификации текста, распознавания именованных сущностей и машинного перевода.

Существует несколько методов и моделей, используемых в NLP и машинном обучении, включая модели глубокого обучения, такие как сверточные нейронные сети, рекуррентные нейронные сети и модели с архитектурой трансформеров. Эти модели можно обучать на больших наборах данных для выполнения сложных задач, таких как языковой перевод, определение смысла текста и ответы на вопросы.

В последние годы *NLP* и машинное обучение приобретают все большее значение в широком спектре приложений, включая чат-боты, виртуальных помощников, обслуживание клиентов, обнаружение мошенничества и медицинскую диагностику. По мере того, как данных становится все больше, а вычислительная мощность растет, возможности *NLP* и машинного обучения становятся шире, и мы можем ожидать еще большего прогресса в будущем.

**II. Обзор литературы по NLP и машинному обучению**

* Основные методы машинного обучения в NLP
  1. Обучение с учителем

Примеры алгоритмов обучения с учителем, используемых в NLP (например, Наивный Байес, Метод опорных векторов, Деревья решений)

Приложения обучения с учителем в NLP (например, анализ тональности, классификация текстов, распознавание именованных сущностей)

* 1. Обучение без учителя

Примеры алгоритмов обучения с учителем, используемых в NLP (например, Наивный Байес, Метод опорных векторов, Деревья решений)

Приложения обучения с учителем в NLP (например, анализ тональности, классификация текстов, распознавание именованных сущностей)

* 1. Обучение с подкреплением

Примеры алгоритмов обучения с подкреплением, используемых в NLP (например, Марковские процессы принятия решений, Q-обучение)

Приложения обучения с подкреплением в NLP (например, диалоговые системы, чат-боты)

* 1. Глубокое обучение

Примеры моделей глубинного обучения, используемых в NLP (например, сверточные нейронные сети, рекуррентные нейронные сети, трансформеры)

Приложения глубинного обучения в NLP (например, машинный перевод, ответы на вопросы, генерация текстов)

* **Техники предобработки текстовых данных**

Методы предварительной обработки текста имеют решающее значение в задачах обработки естественного языка (NLP), поскольку они помогают преобразовывать необработанные текстовые данные в формат, который может быть легко понят и проанализирован моделями машинного обучения. Методы предварительной обработки включают ряд шагов, которые очищают и подготавливают текстовые данные для дальнейшего анализа. Некоторые из наиболее распространенных методов предварительной обработки текста включают токенизацию, удаление стоп-слов, выделение корней и лемматизацию.

Токенизация — это процесс разбиения текста на отдельные слова или токены. Этот шаг важен, поскольку он позволяет модели машинного обучения анализировать текст по одному слову. Удаление стоп-слов включает удаление общих слов, таких как «и», «тот» и «есть», которые не вносят большого вклада в смысл текста. Этот метод помогает уменьшить размерность данных и повысить эффективность модели.

Стемминг — это процесс приведения слова к его корневой форме, который помогает объединить слова, имеющие одинаковое значение, но разные формы. Например, «ходьба», «ходили» и «поход» будут сокращены до «ходить». Лемматизация похожа на стемминг, но включает в себя приведение слова к его базовой форме в зависимости от его части речи. Например, глагол «были» будет сокращен до «быть», а существительное «мыши» будет сокращен до «мышь».

Другие методы предварительной обработки текста включают удаление пунктуации, преобразование всего текста в нижний регистр и обработку специальных символов и числовых данных. Конкретные используемые методы зависят от характера текстовых данных и требований задачи машинного обучения. В целом, предварительная обработка текста является важным шагом в NLP, который может существенно повлиять на точность и производительность моделей машинного обучения.

* **Алгоритмы классификации текстовых данных**
* **Работа с неструктурированными данными**

**III. Обзор литературы по задаче поиска штрихкода по описанию товара**

* Подходы к решению задачи
* Анализ существующих методов и моделей
* Классификация товаров на основе текстовых данных

**IV. Методология работы**

* Выбор архитектуры модели машинного обучения
* Подготовка данных для обучения
* Описание параметров модели
* Обучение и оценка модели

**V. Результаты работы**

* Описание полученной модели машинного обучения
* Сравнение существующих методов и моделей
* Анализ полученных результатов

**VI. Обсуждение и заключение**

* Обсуждение достигнутых результатов
* Возможности дальнейшего улучшения модели
* Заключение работы

VII. Список литературы

# ПРИНЦИП РАБОТЫ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ АРХИТЕКТУРЫ ТРАНСФОРМЕРОВ

## ВЫВОДЫ

ЗАКЛЮЧЕНИЕ

СПИСОК ЛИТЕРАТУРЫ

ПРИЛОЖЕНИЕ 1