Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.8

Ракович
Анастасия Анатольевна

Автоматизация бизнес-процесса обработки типизированных входных данных на базе платформы с применением машинного обучения

АВТОРЕФЕРАТ
на соискание степени магистра информатики и вычислительной техники

по специальности 1-40 81 03 – Искусственный интеллект

Магистрант А.А. Ракович

Научный руководитель
В.И. Романов, кандидат технических наук, доцент

Минск 2019

ВВЕДЕНИЕ

Ежедневно тысячи организаций и их миллионы сотрудников так или иначе занимаются монотонной и однотипной обработкой документов, электронных писем: получение документа, его открытие, извлечение необходимых данных и их занесение в системы, работа с системами. Набор действий, которые сотрудник выполняет с набором документов ежедневно достаточно велик. В большинстве случаев все шаги, которые выполняет сотрудник можно заменить определенными наборами правил, а это значит, что такие действия могут быть полностью или частично автоматизированы.

Реализация эффективного подхода для массовой обработки однотипных документов – хорошая альтернатива, для решения обширного круга задач. Автоматизация этих шагов может позволить сократить время, затрачиваемое на обработку документов, сократить количество возникающих ошибок, появляющихся из-за человеческого фактора, а также сократить количество сотрудников на предприятиях.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования обусловлена потребностью в автоматизации процессов обработки наборов типизированных данных с последующим их использованием.

Цель и задачи исследования

Для достижения данной цели были поставлены следующие задачи:

– реализация компонента для мониторинга почтового ящика;

– определение типа обрабатываемых документов (электронных писем) и получение набора документов для разработки и для тестирования, их обработка и подготовка к использованию;

– подготовка наборов для обучения и тестирования моделей, обучение и тестирование;

– разработка трех моделей: одной классификационной и двух для извлечения данных из документов;

– автоматизация процесса работы с веб-приложением;

– разработка генератора отчета, на основании полученных после тестирования модели данных, сбор и подсчет финальных метрик.

Объектами исследования является методы, подходы и алгоритмы машинного обучения, обработки документов и автоматизации веб-приложений, которые способны решить поставленную задачу.

Предметами работы выступают типизированные данные – электронные письма и веб-приложение, автоматизация которого была выполнена.

Информационная база

В основу информационной базы легли электронные письма, полученные путем мониторинга почтового ящика.

Научная новизна диссертационной работы заключается в создании законченного процесса, позволяющего автоматизировать процесс обработки документов. Были спроектированы, разработаны и обучены модели для бинарной классификации электронных писем и модели для извлечения данных и текстов писем и документов-вложений с применение алгоритмов и методов машинного обучения.

Основные положения, выносимые на защиту

1. Средство автоматизации мониторинга почтового ящика, основанное на проведении контроля новых поступлений с целью последующего выделения типовых документов.
2. Модель бинарной классификации писем, основанная на использовании метода Байеса. Классификация опирается на предварительно проведенное обучение, выполненное с использованием подготовленных наборов типовых документов.
3. Модели для извлечения данных из полученных документов, построенные на основе использования существующих методов искусственного интеллекта. Данные модели основываются на предварительном обучении по заранее подготовленным образцам.
4. Компоненты для обеспечения автоматизации веб-приложения, на основании шагов, выполняемых оператором, обрабатывающим документы.

Теоретическая значимость диссертации заключается в том, что предложены и программно реализованы модели классификации и извлечения

данных из документов, основаны на использовании моделей и методов искусственного интеллекта.

Практическая значимость диссертации состоит в том, что все разработанные компоненты являются самостоятельными модулями и могут быть интегрированы в будущие разработки. Разработанный бизнес-процесс и модуль для генерации отчета для получения метрик для обученных моделей были внедрены в производство, что подтверждено справкой о внедрении.

Структура и объем работы. Структура диссертационной работы обусловлена целью, задачами и логикой исследования. Работа состоит из введения, шести глав, заключения и библиографического списка. Общий объем диссертации составляет 70 страниц. Работа содержит 30 рисунков и 7 листингов. Библиографмческий список включает 30 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении рассмотрено состояние проблемы по обработке больших массивов типизированных данных, дается обоснование актуальности темы диссертационной работы, а также сформулированы основные цели и задачи проводимых исследований.

В первой главе рассматриваются наиболее используемые системы, предоставляющие похожий функционал: их основные возможности, достоинства и недостатки. Сделан вывод о том, что и один аналог не способен предоставить разработанный в диссертационной работе функционал.

Во второй главе был описан процесс проектирования разрабатываемого бизнес-процесса и организация процесса обработки исключительных ситуаций. Также был проведен обзор основный компонентов, которые предоставляет платформа Workfusion, которые планировалось использовать во время разработки и тестирования.

В третьей главе был проведен анализ исходных типизированных данных – электронных писем. Описан реализованный модуль для мониторинга почтового ящика и модуль для получения метаданных письма. Также был описан реализованный бизнес-процесс для применения механизма Optical Character Recognition для документов в формате PDF.

В четвертой главе описан процесс разработки моделей, основанных на методах машинного обучения. Было реализовано, обучено и

протестировано три модели. Одна для бинарной классификации с использованием метода Байеса. Две модели для извлечения данных из документов разного типа с применением методов машинного обучения, предоставляемых компанией Workfusion. Также был описан процесс подготовки и фильтрации документов, которые использовались для тренировки, валидации и тестирования полученных моделей.

В пятой главе была описана реализация компонента, для автоматизации имеющегося веб-приложения. Реализованный компонент является частью общего процесса и позволяет полностью убрать человеческое вмешательство во время заполнения приложения данными и сохранения их, полученными после применения обученных моделей.

В шестой главе были описаны основные метрики, которые использовались для оценки качества обученных моделей. Также описан разработанный модуль для генерации отчета для получения описанных метрик, что позволило заметно сократить время затрачиваемое на оценку качества моделей. Описано проведенное тестирование собранного бизнес-процесса, включающего в себя все разработанные модули.

ЗАКЛЮЧЕНИЕ

В рамках магистерской диссертации был полностью автоматизирован бизнес-процесс для обработки типизированных входных данных, в конкретно рассматриваемом случае это были электронные письма с прикрепленными файлами. В процессе автоматизации были проанализированы, реализованы и применены некоторые методы машинного обучения для решения задачи классификации и извлечения данных из структурированных текстов. Для преобразования PDF документов к виду, пригодному к обработке был реализован и применен механизм Optical Character Recognition. В результате применения описанных выше процессов были получены данные, которые с помощью реализованной автоматизации web-приложения были занесены в необходимую систему.