

Секция «Искусственный интеллект и цифровая трансформация в бизнесе и
государственном управлении»

**Применение быстрого обучения для автоматического определения полей и
классификации документов в ЭДО**

Научный руководитель – Коробова Людмила Анатольевна

Бакланов Олег Владимирович

Выпускник (магистр)

Воронежский государственный университет инженерных технологий, Воронежская
область, Россия

E-mail: ov.baklanov@yandex.ru

Качество распознавания документа и скорость его обработки являются основными параметрами в системе электронного документооборота. Для улучшения этих параметров используются технология FastML и классификация на основе искусственного интеллекта [2].

FastML – это быстрое машинное обучение поиску полей. Для использования модели администратор создает определение документа, внутри которого размещает необходимые для распознавания поля. Далее производит настройку полей. Это может быть установка формата содержимого, правил нормализации поля, очистки от некорректных символов и тд. Затем полям, которые будут обучаться, выставляется настройка «Может иметь регион». Данная функция разрешает модели FastML использовать регион и искать этот регион на странице документа. После этого определение документа сохраняется и публикуется. В разделе обучение поиску полей создается пакет обучения, которые соответствует нашему типу документа. В этот пакет загружаются образцы для обучения модели. Рекомендовано загружать не менее 70 примеров, но чем больше, тем модель лучше обучится. Далее выборка делится на 3 части: 30% - для обучения, 30% - для тестирования и 40% - не используется. На последних будет проверяться работа гибкого описания, созданного моделью. Для обучения модель требуется определить регионы полей и на обучающей выборке, и на тестовой. Это делается для того, чтобы модель сначала обучилась, а потом протестировала себя на тестовых документах. Исходя из результатов теста, модель понимает, какие образцы требуют переопределить регион нахождения поля на странице. Администратор в итоге видит, что на документах, которые модель не смогла определить регион, нужно изменить положение региона и снова запустить обучение модели. Таким образом, спустя несколько итерации обучения модель способна самостоятельно определять регионы полей текущего типа документа. При появлении нового варианта документа, верификатор размечает регионы полей, а документ автоматически передается на обучение и модель адаптируется. Это значительно сокращает время на разработку и поддержку шаблона документа [1].

Классификатор используется для распределения документов по типам на основе различных данных на странице. Это могут быть рисунки, текст, печати, эмблемы компаний и тд. Администратор собирает пакеты по каждому типу документа. На один тип документа рекомендуется не менее 30 образцов. В отличие от модели FastML, классификатор не требует большого количества примеров. Далее документы по каждому типу разделяются на две группы: 60% - для обучения и 40% - для тестирования. Затем администратор заполняет матрицу соответствия классов, где каждому типу документа присваивается класс для его идентификации в системе. Обучение проводится для всех типов документов сразу. Обучение состоит из следующих шагов: анализ графических объектов, анализ текстового сегмента, создание правил соответствий, проверка результатов обучения на тестовой

выборке и автоматическое составление таблицы результата использования модели. В таблице названия строк соответствуют типу документа, а названия столбцов – созданным классам. Если классификатор обучился идеально, то будет заполнена только диагональ матрицы, это значит, что все классы наложились на документы правильно. Бывают и ошибки. В этом случае таблица покажет на какие документы наложился неправильный класс и выделит его красным цветом. Администратор, в свою очередь, добавит некорректно определившийся документ в обучающую выборку и снова запустит обучение. В итоге, классификатор обучится до 100% точности и его можно использовать в системе [3].

Для использования классификатора в системе необходимо добавить его в раздел, занимающийся обработкой поступающих документов, и написать скрипт на языке C#, который будет управлять классификатором во время обработки документов. Это нужно для того, чтобы классификатор работал только на определенном этапе и не мешал системе на других этапах, не требующих использования классификатора. Таким образом при распознавании классификатор понимает к какому типу относится документ, а модель FastML распознает необходимые данные. Основные преимущества использования моделей Классификатора и FastML заключаются в том, что они равнозначны разработке гибкого описания с нуля, при этом занимают меньше времени, что сокращает расходы компании. При появлении новых документов они легко адаптируются к изменениям. Модели требуют минимум поддержки и имеют максимально понятный интерфейс, не требующий специальных навыков разработчика. Кроме того, скорость обработки документа в системе в разы выше, чем при использовании гибких описаний [4]. Использование описанных моделей позволит компании увеличить скорость обработки документов, без потери качества документов, и снизить траты на разработку гибких описаний квалифицированными сотрудниками. Использование технологий FastML и классификации на основе искусственного интеллекта значительно ускоряет процесс обработки документов, обеспечивая высокую точность распознавания и минимизируя ручной труд [5].

Модели помогают автоматизировать не только поиск и классификацию данных, но и адаптацию системы к изменениям. Это делает их незаменимыми инструментами в системах электронного документооборота. FastML является наиболее быстрым инструментом для машинного обучения поиску полей. Это упрощает процесс настройки системы. Администратор может быстро настроить поля для обучения, определить их параметры и предоставить системе примеры для анализа. По мере обучения модель самостоятельно корректирует свою работу, повышая точность распознавания. Классификатор, в свою очередь, отвечает за автоматическое распределение документов по категориям. Он анализирует текстовые данные, графические элементы, печати и другие характеристики, определяет, к какому типу относится документ. Это позволяет быстро сортировать документы без участия человека и снижает вероятность ошибок. В процессе обучения классификатор проходит несколько этапов тестирования, чтобы добиться максимальной точности. Внедрение FastML и классификатора в электронный документооборот обеспечивает надежность и стабильность работы системы, минимизируя влияние человеческого фактора. Использование FastML и классификатора представляет собой прогрессивное решение для оптимизации работы. В результате компания получает мощный инструмент для автоматизации процессов, повышает общую продуктивность и эффективность работы организации.

Источники и литература

- 1) Бутенко Екатерина Александровна, Задорожный Александр Михайлович, Любовинкина Наталья Яковлевна, Потемкина Снежана Владиславовна Сетевое научное издание. 2023. № 1. С. 7-12.URL: <https://sanse.ru/index.php/sanse/article/view/567>

- 2) Тумакова Анна, Интеллектуальная обработка документов. Сферы применения, импортозамещающие решения, комментарии экспертов российского ИТ-рынка.2023. URL: <https://ict-online.ru/analytics/Intellectual-naya-obrabotka-dokumentov-Sfery-priimeneniya-importozameshchayushchiye-resheniya-kommentarii-ekspertov-rossiiskogo-IT-rynka-279905>
- 3) Content AI. URL: <https://contentai.ru/contentreader-engine/klassifikaciya-dokumentov>
- 4) Content AI. Кросс-платформенная версия решения для интеллектуальной обработки информации. URL: <https://softline.ru/about/blog/content-ai-vypustila-kross-platformennuyu-versiyu-resheniya-dlya-intellektualnoj-obrabotki-informacii>
- 5) Content AI и DocTrix: новые возможности для оптимизации рабочих процессов. URL: <https://softline.ru/about/blog/content-ai-i-doctrix-novye-vozmozhnosti-dlya-optimizacii-rabochih-processov>