Скриншот инструмента маркировки Form Recognizer

Как создать считыватель удостоверений личности и паспортов менее чем за час с помощью Azure AI

Мик Влишоувер
Технический специалист по искусственному интеллекту @ Microsoft

10 ноября 2020 г.

В ряде отраслей общепринято и требуется законом обрабатывать идентификационные данные пользователей с использованием действительного подтверждения личности. Государственные учреждения, банковский и страховой сектор, а также работодатели являются примерами организаций, которым разрешено обрабатывать копию вашего подтверждения личности в Нидерландах (**источник**). Для того чтобы оцифровать эти идентификационные данные, сотруднику часто приходится перепечатывать эти данные в компьютерной системе.

Это звучит не очень эффективно... Представьте себе выполнение этой повторяющейся и подверженной ошибкам задачи в огромных объемах, например, в агентстве по трудоустройству. Что, если мы сможем использовать искусственный интеллект для извлечения данных из доказательства личности напрямую? В этой статье я объясню, как построить модель ИИ для этой цели менее чем за час, не требуя глубоких знаний в области науки о данных.

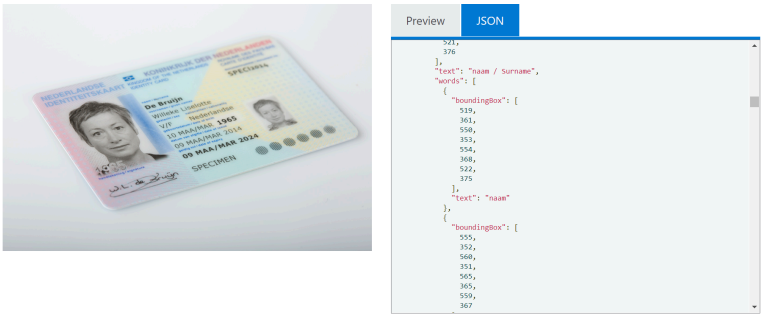
> **Принципы «Знай своего клиента» (KYC) в сфере финансовых услуг требуют, чтобы специалисты прилагали усилия для проверки личности, пригодности и рисков, связанных с поддержанием деловых отношений.**

> Источник: **Википедия**

Введение в Распознаватель форм.

В прошлом эту проблему пытались решить с помощью **оптического распознавания символов** (OCR). Проблема использования OCR для таких задач заключается в том, что в конечном итоге вы получаете весь распознанный текст с указанием его положения в документе. Как

разработчик/аналитик, вы несете ответственность за интерпретацию этих данных и превращение их в структурированные данные. Это непростая задача, поскольку положения зависят от масштаба и угла захваченного документа, поэтому для улучшения механизма OCR часто требуется использование алгоритмов машинного обучения.



Пример результата OCR, включая позиции (ограничивающие рамки)

Azure Form Recognizer — это когнитивная служба, которая позволяет вам создавать автоматизированное программное обеспечение для обработки данных с использованием технологии машинного обучения. Определите и извлеките текст, пары «ключ/значение» и табличные данные из ваших документов форм — служба выводит структурированные данные, которые включают связи в исходном файле. Вы быстро получаете точные результаты, которые соответствуют вашему конкретному контенту без серьезного ручного вмешательства или обширных знаний в области науки о данных (**источник**).

Form Recognizer **состоит из нескольких служб** , и сегодня мы будем использовать службу «Пользовательские модели» для обучения нашей собственной модели ИИ с использованием маркированных данных. Учетная запись Azure является предварительным условием для этого пошагового руководства и может быть получена через **azure.com/free** .

1. Создайте набор обучающих данных

Чтобы начать работу с пользовательскими моделями Form Recognizer, вам нужно всего лишь пять образцов форм ввода. Сбор образцов удостоверений личности и паспортов может быть обременительным, поскольку он включает в себя чрезвычайно конфиденциальные данные PII, но есть способы обойти это. Например, правительство Нидерландов предоставило **бесплатный набор данных официальных голландских документов** . Загрузите образцы и разделите их на две папки на своем ПК: одну для удостоверений личности и одну для паспортов. Это набор данных, который мы будем использовать для обучения нашей первой модели.

Создать ресурс Form Recognizer и хранилище Blob-объектов

Если вы используете Form Recognizer впервые, вам нужно будет развернуть два ресурса в Azure. Ресурс Form Recognizer, который

обеспечивает доступ к API, и учетная запись хранилища Blob, которая будет использоваться для безопасного хранения ваших данных обучения в вашей собственной подписке. Следующие шаги немного административные, но большинство из них вам нужно будет выполнить только для вашего первого проекта.

1. **Создать ресурс распознавания форм**
2. **Создайте учетную запись хранения (BLOB-объектов)** со всеми настройками по умолчанию.
3. **Настройте кросс-доменный обмен ресурсами в вашем хранилище Blob-объектов**
4. **Создайте контейнер и загрузите данные о ваших тренировках через Azure Storage Explorer** . Убедитесь, что вы поместили удостоверения личности и паспорта в разные папки.

2. Маркируйте и обучайте модель

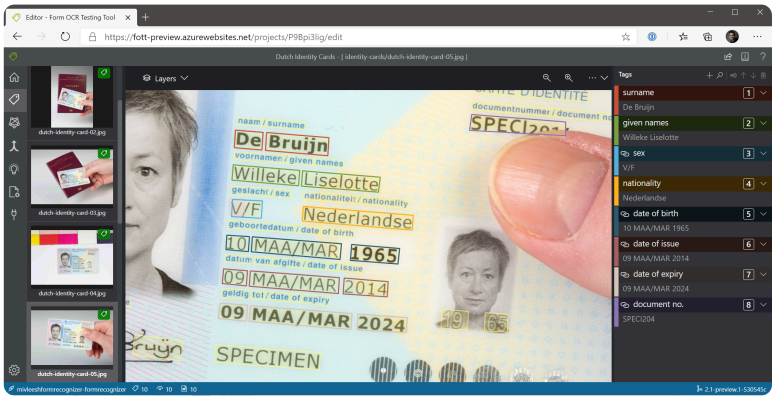
Теперь, когда все создано, мы готовы начать процесс маркировки. Инструмент маркировки Form Recognizer (он же FOTT) — это веб-решение, которое можно посетить через **fott-preview.azurewebsites.net** , или вы можете **использовать образы Docker для его размещения самостоятельно** . Инструмент маркировки OCR Form также доступен как **проект с открытым исходным кодом на GitHub** , что позволяет вам интегрировать инструмент маркировки в ваше собственное приложение.

Откройте инструмент маркировки и выполните следующие шаги:

1. **Подключите свою учетную запись хранения к инструменту маркировки**
2. **Создайте новый проект** . Выберите путь к папке, в которой вы сохранили свои удостоверения личности на предыдущем шаге.

Начните маркировать свои формы

Далее вы создадите теги (метки) и примените их к текстовым элементам, которые вы хотите, чтобы модель распознавала. В нашем примере мы создадим теги для всех элементов на удостоверении личности. Инструмент маркировки имеет несколько удобных сочетаний клавиш для ускорения маркировки. Щелкните по желтому полю вокруг немаркированной части вашего текста и нажмите цифру/клавишу, которая видна на теге. В нашем примере ниже это означает, что вы можете выбрать « *De Bruijn* » и нажать « **1** », чтобы назначить выбор « *фамилии* «.



Повторите эти шаги для всех документов в нашем наборе данных. Закончили? Теперь пора начать обучение нашей модели. Щелкните значок «Обучение» (третий сверху), чтобы открыть страницу обучения. Затем щелкните кнопку **«Обучение»** , чтобы начать обучение модели.

После завершения обучения проверьте значение **Average Accuracy** . Если оно низкое, вам следует добавить больше входных документов и

Train Result

Model ID: ffd2063c-d32e-42cb-b88e-7ab22122bf88

Tag	Estimated Accuracy
date of birth	0.6
date of expiry	0.7
date of issue	0.7
document no.	1
given names	0.8
nationality	1
sex	1
surname	0.9

повторить шаги выше. Документы, которые вы уже пометили, останутся в индексе проекта.

Наши данные обучения имеют одинаковые значения на всех удостоверениях личности, поэтому точность не будет супер. Вы можете легко повысить точность, пометив примеры удостоверений личности разными значениями.

3. Оцените свою модель

После того, как вы обучили свою модель, пришло время оценить ее. Нажмите на значок Predict (лампочка) слева, чтобы протестировать свою модель. Загрузите документ, который вы не использовали в процессе обучения, и спрогнозируйте значения, используя свою модель ИИ.

В зависимости от сообщенной точности, вы можете захотеть провести дополнительное обучение, чтобы улучшить модель. После того, как вы сделали прогноз, проверьте значения достоверности для каждого из примененных тегов. Если среднее значение точности обучения было высоким, но оценки достоверности низкие (или результаты неточные), вы должны добавить файл, используемый для прогнозирования, в обучающий набор, пометить его и снова обучить.

Следующие шаги

Следующим шагом будет интеграция модели в ваше приложение. Form Recognizer предлагает простой в использовании REST API вместе с различными SDK (.NET, Python, Java и JavaScript) для облегчения интеграции.

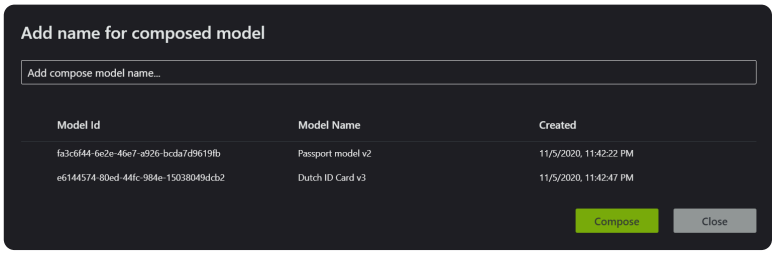
В конце концов, текущая модель может быть расширена водительскими правами, другими европейскими документами,

удостоверяющими личность, или голландскими документами иностранцев, например. Просто создайте новую модель, как вы сделали для удостоверения личности, и объедините их позже с помощью **model compose** . Другими дополнениями могут быть добавление проверки подлинности или скрипт для получения фотографии и подписи.

Будьте готовы к непрерывному циклу обучения, особенно в начале проекта. Вероятно, вам нужны данные, которые лучше подходят для вашего сценария, а не высококачественные студийные фотографии, которые мы использовали. В этом сценарии вам, вероятно, понадобятся удостоверения личности с водяными знаками и сканы низкого качества в ваших обучающих данных.

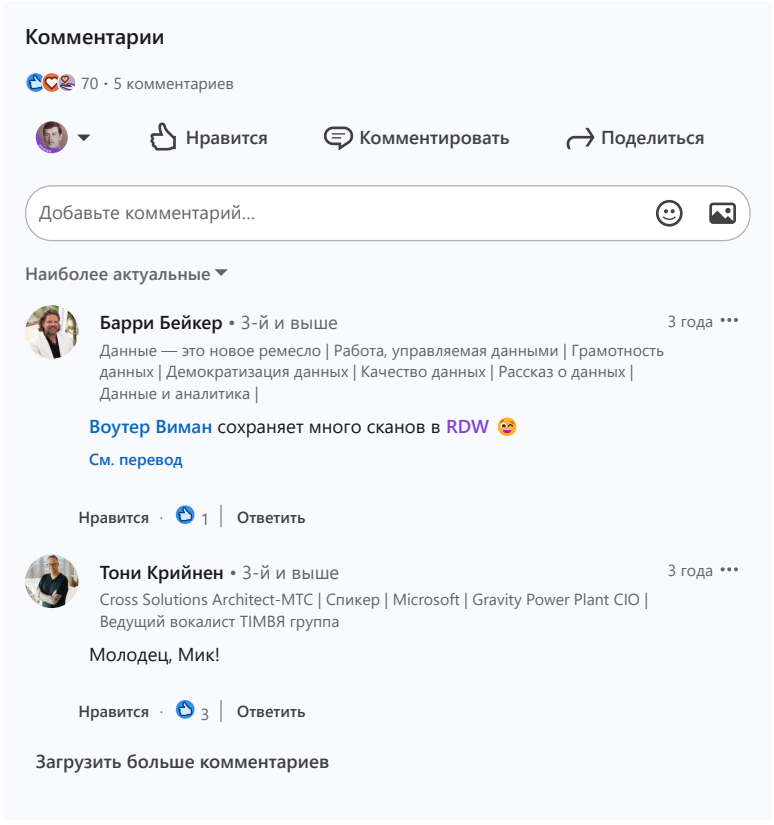
Надеюсь, эта статья вдохновит вас начать работу с Form Recognizer и обучить модель по вашему выбору. (И надеюсь, что вскоре системы не будут требовать от вас ввода ваших данных И хранения копии подтверждения личности)

Не стесняйтесь обращаться, если у вас есть вопросы или замечания. Я с радостью помогу вам открыть для себя возможности ИИ на Azure.



Объединение нескольких предметно-ориентированных моделей в одну главную модель (составная модель)

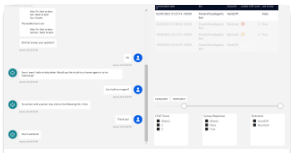
Пожаловаться на эту статью



Мик Влишоувер
Технический специалист по искусственному интеллекту @ Microsoft

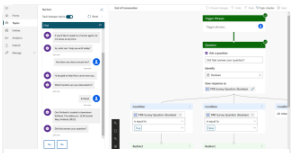
+ Отслеживать

Другие статьи Мика Влишоувера



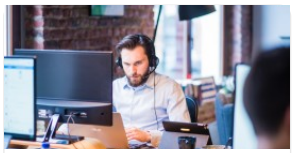
Улучшите ботов Power Virtual Agent с помощью Power BI и...

Мик Влишоувер в LinkedIn



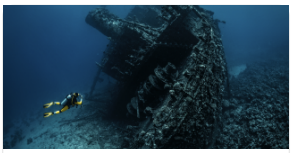
Преодолейте ограничения Power Virtual Agents с помощью Bot Framework...

Мик Влишоувер в LinkedIn



Обслуживание клиентов во время кризиса. Часть II. Объединение ботов и...

Мик Влишоувер в LinkedIn



Добыча знаний. Объяснение модного слова.

Мик Влишоувер в LinkedIn

См. все статьи (5)