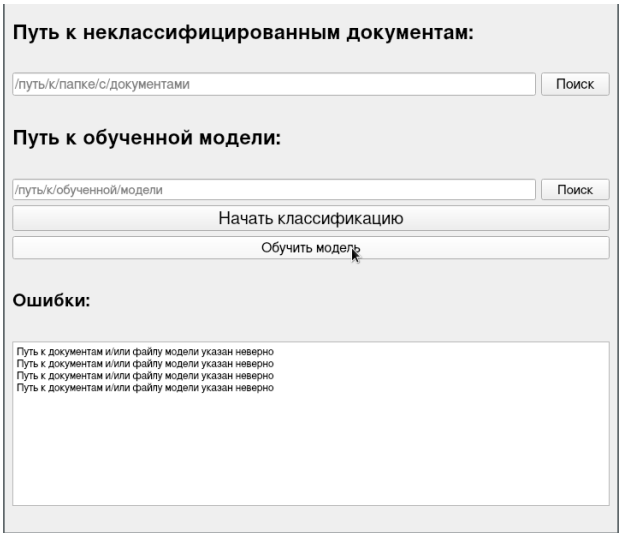
Для чего нужно приложение docthrush

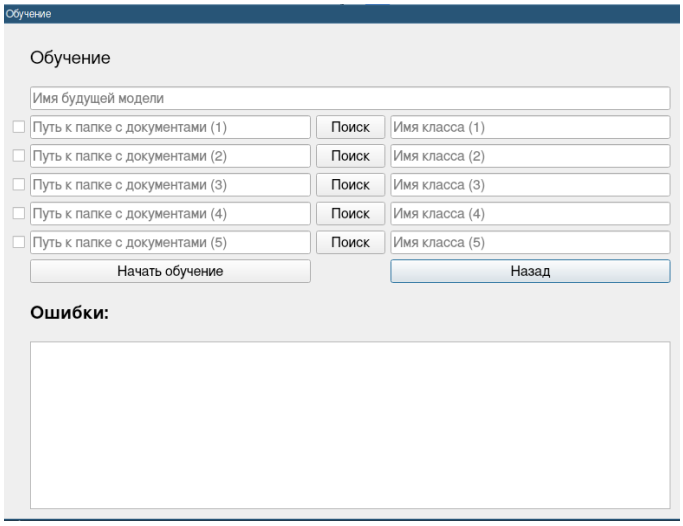
Приложение docthrush позволяет классифицировать графические документы (файлы форматов pdf/png/jpg и др.) без необходимости просматривать их содержимое. Программа при помощи технологии нейронных сетей научится распознавать типы документов и автоматически распределять их по соответствующим папкам, без ручного вмешательства.

Как использовать приложение docthrush

После установки, по умолчанию на рабочем столе появится ярлык docthrush. Его нужно открыть. Чтобы начать использовать приложение, нужно обучить модель. Модель - файл нейронной сети, оканчивающийся на “.h5”. Для того, чтобы начать обучать модель, нужно выбрать соответствующий пункт в главном меню приложения - “Обучить модель”.

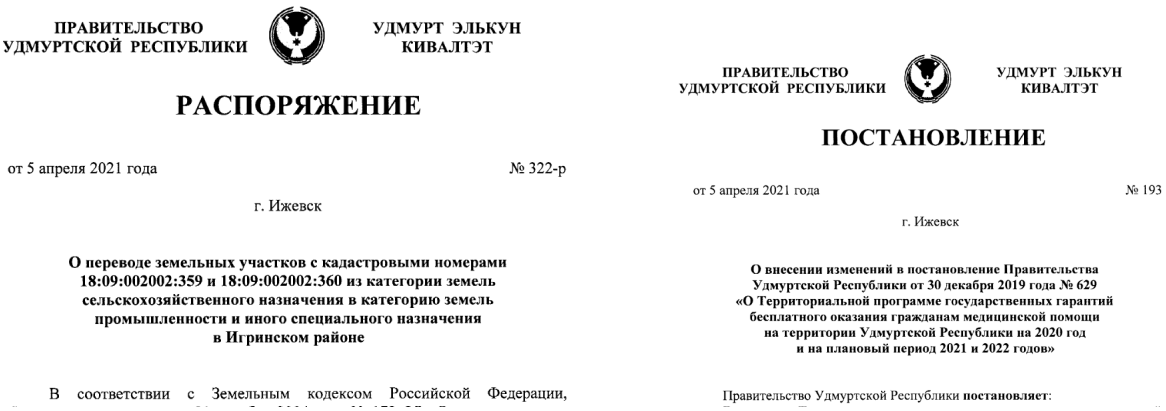


После нажатия на кнопку, на экране появится окно обучения модели:



Подготовка файлов к обучению

Чтобы обучить модель, нужно сначала вручную распределить некоторую часть файлов. Предположим у вас есть два типа различных файлов - “распоряжения” и “постановления”, выглядят они примерно одинаково, отличаются у них только заголовки:



ПРАВИТЕЛЬСТВО
УДМУРТСКОЙ РЕСПУБЛИКИ

УДМУРТ ЭЛЬКУН
КИВАЛТЭТ

РАСПОРЯЖЕНИЕ

от 5 апреля 2021 года № 322-р

г. Ижевск

О переводе земельных участков с кадастровыми номерами 18:09:002002:359 и 18:09:002002:360 из категории земель сельскохозяйственного назначения в категорию земель промышленности и иного специального назначения в Игринском районе

В соответствии с Земельным кодексом Российской Федерации,

ПРАВИТЕЛЬСТВО
УДМУРТСКОЙ РЕСПУБЛИКИ

УДМУРТ ЭЛЬКУН
КИВАЛТЭТ

ПОСТАНОВЛЕНИЕ

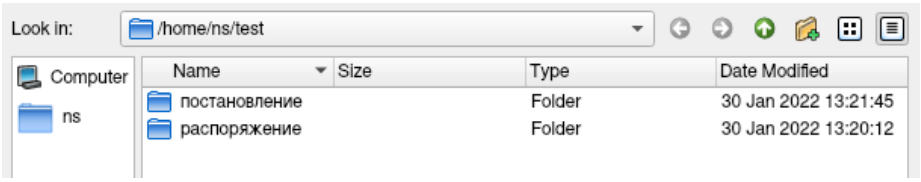
от 5 апреля 2021 года № 193

г. Ижевск

О внесении изменений в постановление Правительства Удмуртской Республики от 30 декабря 2019 года № 629 «О Территориальной программе государственных гарантий бесплатного оказания гражданам медицинской помощи на территории Удмуртской Республики на 2020 год и на плановый период 2021 и 2022 годов»

Правительство Удмуртской Республики **постановляет**:

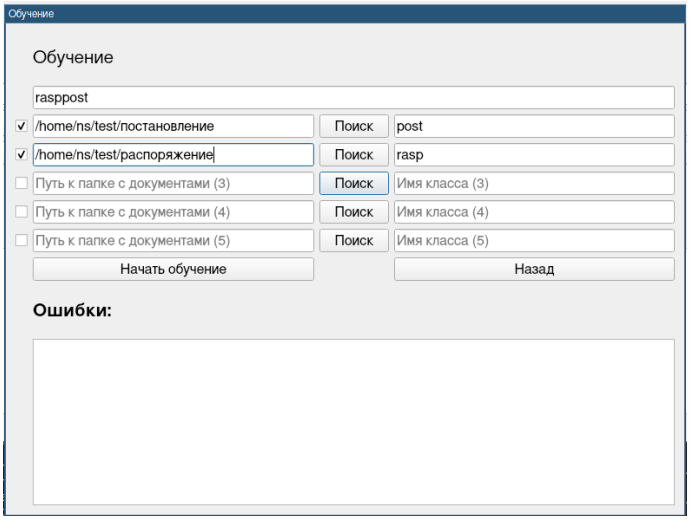
Расположите некоторое количество (в зависимости от сложности документов, для простых, вроде указанных в примере, достаточно использовать 4-5 штук на каждую категорию) одинаковые документы с заголовком “распоряжение” в папку на диске, которая будет называться “распоряжения”, а с заголовком “постановление в папку “постановления”. Папки можно создать где угодно, главное - запомнить их расположение на диске. Таким образом у вас должно получиться 2 папки, в одной из которых лежат только документы распоряжения, а во второй - только постановления.



Процесс обучения

После подготовки файлов, возвращаемся в окно “обучение” приложения docthrush. В поле “имя будущей модели” введите название. Вводить желательно что-нибудь на английском, а само название не принципиально и на работу повлиять не должно.
Далее выделяем галочки. Выделить нужно столько, сколько у вас категорий. В нашем случае категории всего две - постановление и распоряжение -, поэтому выделяем две галочки.
Справа каждой выделенной галочки нужно указать путь к папке с документами. Для того, чтобы найти путь, достаточно нажать кнопку поиск и выделить папку. В нашем случае выделяем те две папки, которые мы подготовили в предыдущем пункте.

Каждому классу документов нужно ввести название; в нашем случае используем “post” и “rasp” (названия лучше вводить на английском). Так будут называться папки, в которые нейронная сеть будет распределять файлы.
После указания всех пунктов окно обучения должно выглядеть похожим образом:



После того, как вы убедились, что все папки и названия указаны правильно, можно нажимать кнопку “начать обучение”. Программа не будет отвечать, но через некоторе время, она завершит работу.

Получение модели

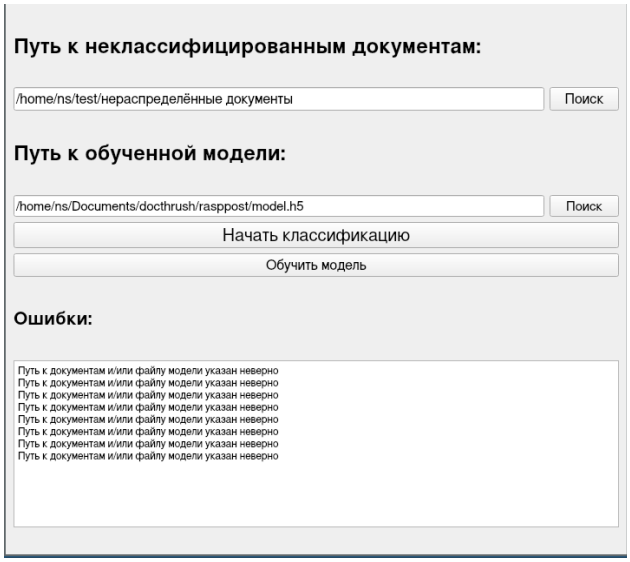
После завершения обучения, полученная модель будет сохранена в папку “Документы”, на вашем компьютере, в подпапку docthrush. В папке docthrush вы найдёте папку с названием, которое вы указали в окне “имя будущей модели”, в нашем случае - папка rasppost. Внутри неё лежит модель, с расширением “.h5”.

Подготовка файлов к распределению

Перед тем, как программа начнёт свою работу, сложите все нераспределённые документы в одну папку.

Использование обученной модели

Возвращаемся к главному окну приложения, нажав кнопку “назад” или просто закрыв окно обучения. В поле путь к неклассифицированным документам указываем путь к папке, которую мы организовали в прошлом пункте. Чтобы просто выбрать папку, нажмите на кнопку “Поиск”.
В поле “путь к обученной модели” указываем путь к модели с расширением “.h5”. Если вы никуда её не переместили, то она всё так же находится по пути документов в папке docthrush, в подпапке с названием вашей модели.
После указания путей, окно должно выглядеть похожим образом:



Если вы уверены, что все пути указаны верно, то можете нажать “Начать классификацию”, тогда все документы в выбранной папке автоматически распределятся, согласно обученной модели.