КТО НЕ ИДЁТ ВПЕРЕД, ТОТ ИДЁТ НАЗАД

DR NATASHA KHRAMTSOVSKY

ОБО МНЕ: НАТАША ХРАМЦОВСКАЯ / DR NATASHA KHRAMTSOVSKY

ПРОСМОТРЕТЬ ПРОФИЛЬ

Вы можете скачать или посмотреть список моих публикаций (с активными гиперссылками на большинство из них) на март 2015 г., посмотреть мои презентации и документы и мои аудио- и видеозаписи на YouTube

You can also download or view a list of my publications (with hyperlinks to full texts provided where possible) as of March 2015, browse my

ПОНЕДЕЛЬНИК, 26 МАРТА 2018 Г.

Машины читают архивные документы: Программное обеспечение для распознавания рукописного текста

Данная статья д-ра Ричарда Данли (Dr Richard Dunley – на фото) была опубликована 19 марта 2018 года на блоге Национальных Архивов Великобритании.

Любой исследователь, который пользовался онлайн-архивами газет, хранилищами оцифрованных книг или даже такими ресурсами Национальных Архивов, как «Бумаги Кабинета министров онлайн» (Cabinet Papers Online, http://www.nationalarchives.gov.uk/cabinetpapers/) понимает, какую революцию произвела технология оптического распознавания символов (OCR).

Именно эта технология позволяет нам искать не только по названию и дате, но и непосредственно по словам, написанным внутри книги, газеты или архивного документа. Технология OCR изменила способы, при помощи которых многие учёные проводят свои исследования и открыла новые, немыслимые ранее огромные пространства для научных исследований.

Для тех из нас, кто работает с архивными коллекциями, эта революция всегда сопровождалась оговоркой - OCR не работает для рукописных документов. Именно по этой причине у нас вызвала такой живой интерес новая платформа Transkribus (https://read.transkribus.eu/transkribus/), разработанная в рамках финансируемого Евросоюзом проекта READ (https://read.transkribus.eu/). Впервые появилась потенциальная возможность использования компьютеров для «чтения» рукописных документов.

documents and Powerpoint presentations and watch my YouTube channel

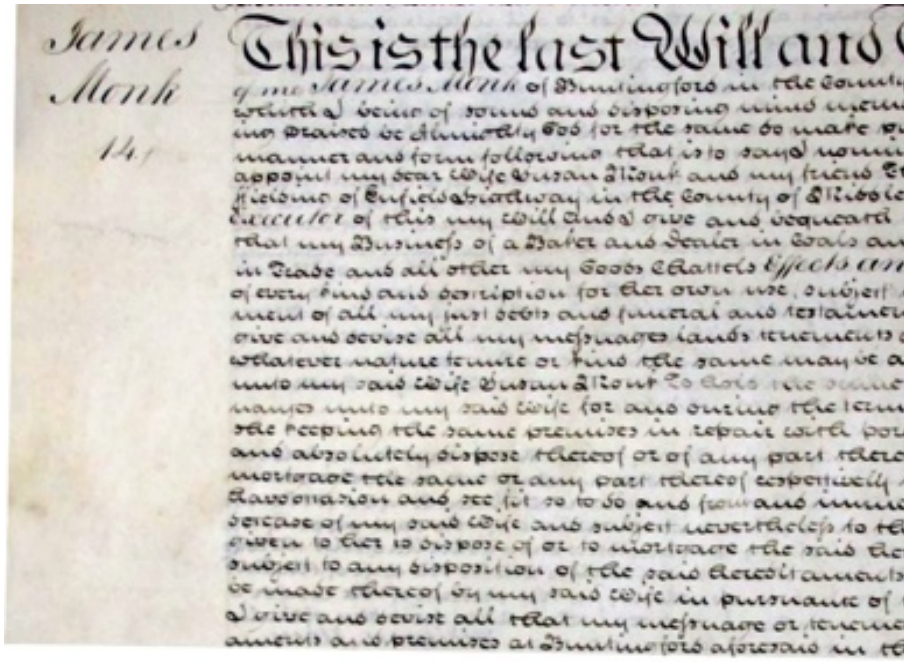
ПОИСК ПО БЛОГУ / SEARCH THIS BLOG

Поиск

ИСКАТЬ ПО БЛОГУ / SEARCH THIS BLOG 2

search...

АРХИВ БЛОГА / BLOG ARCHIVES



Документ PROB 11/2105/1: Можете ли Вы прочитать это завещание, написанное в 1849 году?

Технология, на которую опирается решение Transkribus, пока что очень новая, и Национальные Архивы проводит пилотный проект с тем, чтобы протестировать возможность применения этого программного обеспечения для распознавания рукописного текста (handwritten text recognition, HTR).

В этом проекте мы решили использовать материалы из нашей коллекции завещаний PROB 11 (http://discovery.nationalarchives.gov.uk/details/r/C12122). Причины этого выбора были во многом обусловлены технологическими вопросами - в этих томах содержатся сделанные клерками копии завещаний, поэтому стиль почерков очень однородный. Завещания являются юридическими документами и, как следствие, их тексты структурированы и содержат шаблонные фразы и выражения. Это также исключительно интересная коллекция документов, содержащих сведения о людях, местах, материальных благах, социальных и экономических взаимоотношениях и иных факторах, во времени и в пространстве. Однако, как может подтвердить любой, кто использовал эти документы, это не самое легкое чтение – и по этой причине они представляются отличным тестовым материалом для новой технологии.

Программное обеспечение Transkribus работает путем обучения модели на точных текстах документов. Исследователи загружают графические образы ряда своих документов, а затем сопоставляют правильную транскрипцию с текстом в изображениях. Это позволяет модели изучить стиль почерка и шаблоны словоупотребления. Эти обучающие данные называются «опорной истиной» (ground truth). Обученная модель затем может быть использована для автоматического транскрибирования документов, похожих с точки зрения языка, почерка и т.д. Как можно ожидать, чем больше обучающих данных Вы загружаете, тем лучше результаты, которые Вы можете получить от своей модели.

С 1 января 2019 года конкурсы, запросы котировок и...

Технология блокчейна находится на курсе, ведущем к...

Масштабные поправки в закон о контрактной системе:...

Горячая пора: Пересмотренный стандарт ISO 15489 и ...

Расширяется перечень проверок, информация о которы...

Горячая пора: Пересмотренный стандарт ISO 15489 и ...

Стандартизация по-нашему: Международный стандарт I...

Судебная практика: Обращение граждан и защита чест...

Судебная практика: Начальник отдела Роспотребнадзо...

Проект «Основных принципов для хранилищ, выступающ...

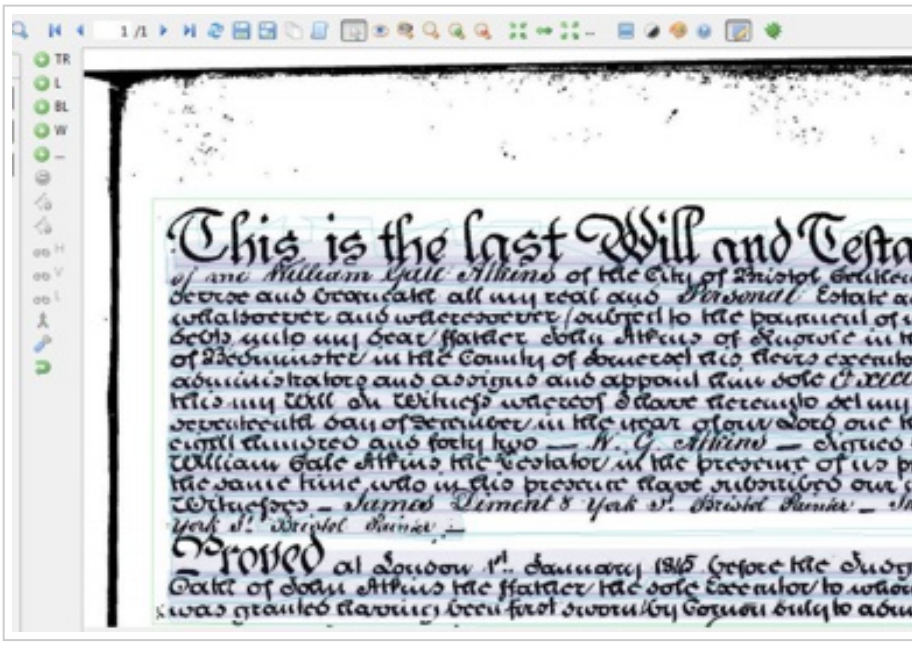
Международный автоматический обмен информацией и д...

Архивные носители информации – не самый процветающ...

Росархив об удаленном использовании архивных докум...

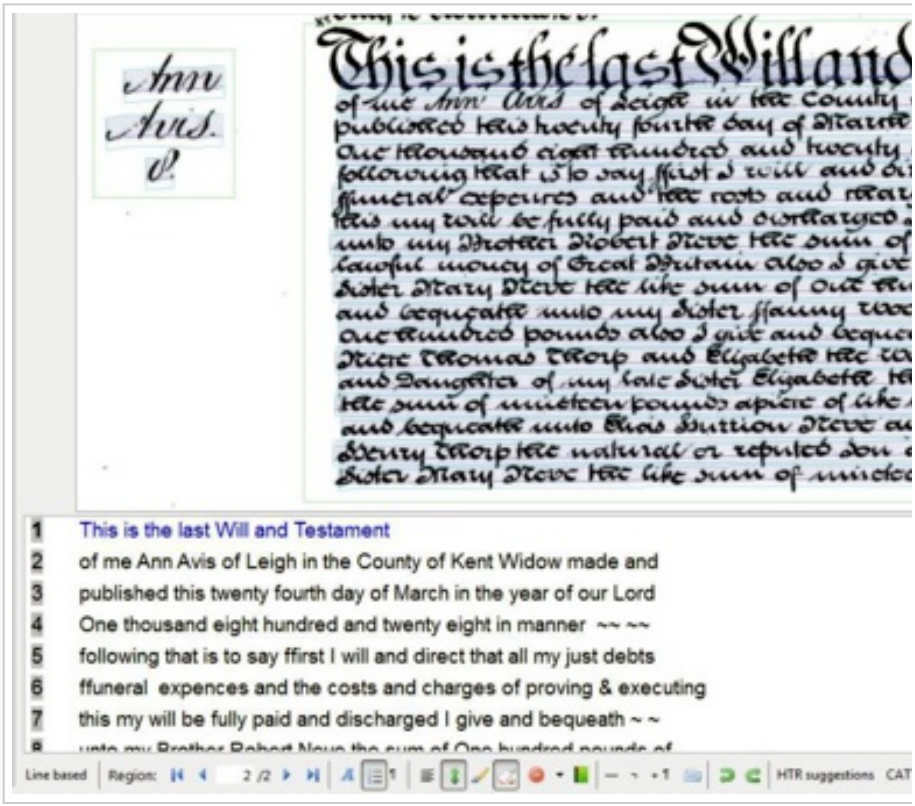
Журнал государственных

Мой комментарий: *На самом деле последнее утверждение неверное – точнее, его можно принять с очень существенной оговоркой: обучающие данные должны быть тщательно подобраны и, в идеале, не противоречить друг другу. Процесс формирования обучающей выборки – исключительно трудоемкая и ответственная работа, требующая высокой квалификации. Попытка же бездумно загрузить первые попавшиеся образцы гарантирует неудачу.*



Сегментация – выделяются области текста и базовые линии

Первым этапом HTR-процесса является загрузка изображений Ваших документов на платформу, а затем выполнение задачи, называемой «сегментацией». При её выполнении выделяются области и строки текста. Результаты сегментации говорят программному обеспечению, где искать текст. Данный процесс в основном автоматизирован, но иногда необходимо проверить и скорректировать его результаты. Как только сегментация завершена, Вы можете либо загрузить свои обучающие данные, либо - когда у Вас уже есть обученная модель - запустить программное обеспечение HTR для создания автоматической транскрипции документа.



Ввод обучающих данных – правильной транскрипции

Мы начали экспериментировать с программным обеспечением некоторое время тому назад и получили неплохие результаты на модели, обученной на относительно небольшом наборе обучающих данных (примерно 15 тысяч слов). Точность OCR и HTR-распознавания оценивается по таким показателям, как процент ошибочных слов (Word Error Rate, WER) и процент ошибочных символов (Character Error Rate, CER). Наша первая модель достигла WER=39% и CCB=21%. Воодушевленные этим, мы подготовили расширенный набор обучающих данных (примерно 37 тысяч слов) и новую модель. К нашей радости, это привело к существенному повышению качества распознавания, с WER=28% и CER =14%.

Сравнение точного текста с результатом автоматического транскрибирования

Это был неплохой результат, однако ясно было, что если более четверти

документации по м...

ГОСТы на сайте Росстандарта: Что почитать?

Вопросы нормативного регулирования в программе «Ци...

Судебная практика: Сотрудник был наказан за несвое...

Арбитражная практика: Электронные товарные накладн...

США: Национальный институт стандартов и технологий...

США, штат Техас: Архивно-библиотечная служба объяс...

Удаленная биометрическая идентификация физических ...

► февраля (50)

► января (50)

► 2017 (702)

► 2016 (716)

► 2015 (866)

► 2014 (819)

► 2013 (726)

► 2012 (744)

► 2011 (690)

► 2010 (662)

► 2009 (212)

► 2008 (238)

► 2007 (12)

► 2006 (7)

► 2005 (6)

► 2004 (1)

всех слов оказались неправильными, то нужно приложить дополнительные усилия. Проблема заключалась в том, что ручное транскрибирование больших количеств этих завещаний - сложная и трудоёмкая операция. По этой причине мы обратились к нашему сообществу онлайн-добровольцев с просьбой помочь сформировать ещё более обширный набор обучающих данных. Благодаря замечательной работе многих увлечённых людей мы быстро накопили еще 60 тысяч слов обучающих данных, которые в настоящее время используются для обучения новой улучшенной модели.

Мы возлагаем большие надежды на эту новую модель, но справедливости ради необходимо сказать, что пройдёт ещё некоторое время, прежде чем мы сможем положиться исключительно на компьютеры для того, чтобы те читали для нас все эти непростые рукописные документы. Пока же технология этого типа предлагает другие потенциальные возможности, в первую очередь в плане поиска по ключевым словам, которые могут оказать трансформирующее влияние на использование архивных коллекций уже в краткосрочной перспективе. Проще говоря, Вы можете использовать эту технологию для поиска по тексту рукописных документов, даже если уровень точности недостаточно хорош для создания их транскрипции. Это связано с тем, что транскрипция может отобразить только один вариант для каждого слова на странице, в то время, как само программное обеспечение генерирует множество возможных вариантов для каждого слова. Используя умные инструменты, Вы можете искать с учетом этих вариантов и находить правильное слово с гораздо большей вероятностью.

Данная технология способна революционизировать работу исследователей с архивными коллекциями, и мы с огромным интересом экспериментируем с ней. Однако эта работа возможна только благодаря увлечённости и самоотверженности наших добровольцев, которые выполнили основную часть трудоемкой работы по транскрибированию. Это позволяет ещё раз подчеркнуть взаимосвязь между восхитительными новыми цифровыми технологиями и более традиционной архивной практикой.

Мы намерены продолжить работу по освоению HTR-технологии и вскоре сообщим о новых результатах, полученных с использованием нашей новой модели.

Ричард Данли (Richard Dunley)

Источник: блог Национальных Архивов Великобритании
http://blog.nationalarchives.gov.uk/blog/machines-reading-the-archive-handwritten-text-recognition-software/

АВТОР: НАТАША ХРАМЦОВСКАЯ НА 11:00
ЯРЛЫКИ: АРХИВНОЕ ДЕЛО, АРХИВНЫЕ ТЕХНОЛОГИИ, ВЕЛИКОБРИТАНИЯ, НАЦИОНАЛЬНЫЕ АРХИВЫ, ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

КОММЕНТАРИЕВ НЕТ:

ОТПРАВИТЬ КОММЕНТАРИЙ