??????????????????????

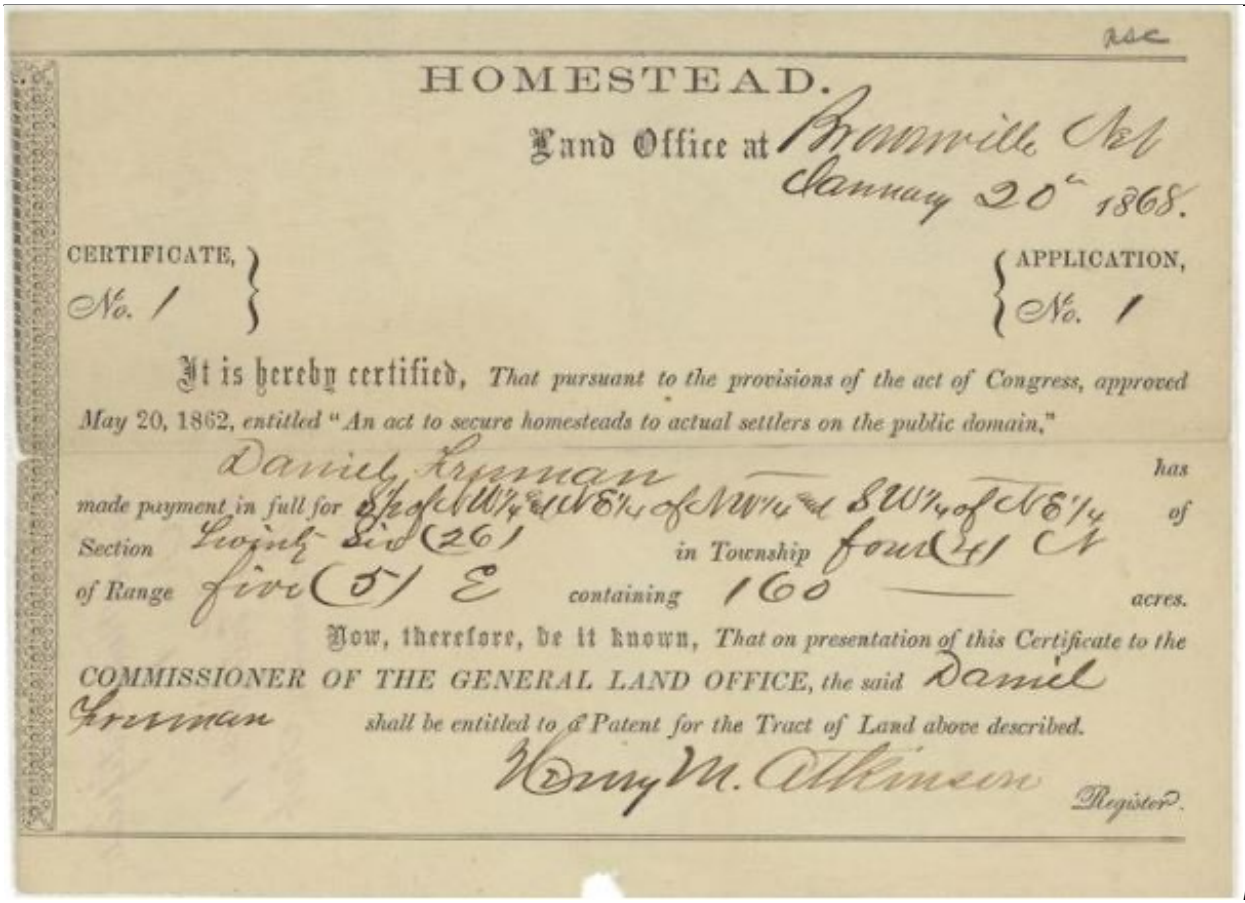
??????????????????????

???Fan Zhang



The American real estate industry is under pressure to leverage big data and evidence-based approaches. The growth of many successful artificial intelligence (AI) real estate startups such as Zillow, Compass, and others are a testament to the value that AI can bring to an enterprising company in this industry. However, the U.S. legal environment provides unique challenges to leveraging AI for real estate.

The Problem with U.S. Deeds



HOMESTEAD.

Land Office at Brownville Neb
January 20th 1868.

CERTIFICATE, No. 1

APPLICATION, No. 1

It is hereby certified, That pursuant to the provisions of the act of Congress, approved May 20, 1862, entitled "An act to secure homesteads to actual settlers on the public domain,"

Daniel Freeman has made payment in full for Lots 3 & 4 & SE¼ of NW¼ & SW¼ of NE¼ of Section Twenty Six (26) in Township four (4) North of Range five (5) East containing 160 acres.

Now, therefore, be it known, That on presentation of this Certificate to the COMMISSIONER OF THE GENERAL LAND OFFICE, the said Daniel Freeman shall be entitled to a Patent for the Tract of Land above described.

Henry M. Atkinson Register.

Image credit to the [National Archives Catalog]

The above image consists of Homestead Entry Number 1, Brownsfield, Nebraska, Land Office, dated January 20, 1868. This is a historical land record held in the National Archives, which confers ownership of a homestead to the first Nebraskan homesteader Daniel Freeman. But in the United States, unlike many any other countries, there is no centralized database that proves a trusted chain of ownership and transfers of properties and rights to enjoyment between legal entities. Every county, parish, or district might have different forms and customs with what amounts to a claim of ownership over a piece of property, land, and right of enjoyment over said property or land. Moreover, in New England states, they are typically recorded at the town level.

The vast majority of American land records exist in the form of deeds, which are records of property and rights transfers between individuals and corporations. Although publicly available archives of deeds exist, those archives are neither thorough nor up to date. For example, the General Land Office Records maintained by the Bureau of Land Management is a data base of 2 million pre-1908 land records for 30 “public land states,” but the data base does not include records for any of the original 13 colony states nor any records after 1908. Essex County, Pennsylvania, is unusual in that it maintains digitized land records from historical to present, but most counties do not.

This flawed approach hampers the entire real estate industry. For example, suppose a bank wants to know the answers to relevant legal proof-of-ownership questions such as “who is the owner,” “when was the property transferred,” “how much is owed,” and “who is the insurer.” Finding those answers means research millions of real estate documents, many of which are undigitized and unlabeled.

Currently, many claims are processed as-is, without any prior digitization or automatic processing, requiring legions of lawyers and clerks to scrutinize deeds and perform data entry to keep track of which rights and properties belongs to whom.

Can we do better?

Exhibit 5.12.7-1 Mutual Collection Assistance Request (MCAR) NFTL

Form 668(Y) (Rev. February 2004)	Department of the Treasury - Internal Revenue Service **Notice of Federal Tax Lien**	
Area:	Serial Number: *NNNNNNNN*	*For Optional Use by Recording Office*

As provided by sections 6321, 6322, and 6323 of the Internal Revenue Code, we are giving a notice that taxes (including interest and penalties) have been assessed against the following-named taxpayer. We have made a demand for payment of this liability, but it remains unpaid. Therefore, there is a lien in favor of the United States on all property and rights to property belonging to this taxpayer for the amount of these taxes, and additional penalties, interest, and costs that may accrue.

Name of Taxpayer: *First Name Line*
Second Name Line

Address: *Street Address*
City, State ZIP

IMPORTANT RELEASE INFORMATION: For each assessment listed below, unless notice of lien is refiled by the date given in column (e), this notice shall, on the day following such date, operate as a certificate of release as defined in IRC 6325(a).

Kind of Tax (a)	Tax Period Ending (b)	Identifying Number (c)	Date of Assessment (d)	Last Day for Refiling (e)	Unpaid Balance of Assessment (f)
Income	*MM/DD/YYYY*	*XXX-XX-NNNN*	*MM/DD/YYYY*	*MM/DD/YYYY*	*NNN,NNN.NN*

****This amount is due, owing, and unpaid to the government of *[Treaty Partner]* and is being collected on behalf of *[Treaty Partner]* under the provisions of Article *[NN]* of the United States – *[Treaty Partner]* Income Tax Convention and applicable provisions of the Internal Revenue Laws of the United States****

Place of Filing: *Recording Official*
Recording Office
City, State

Total $ *NNN,NNN.NN*

This notice was prepared and signed at *City, State*, on this,

the *DD* day of *Month*, *YYYY*.

Signature: *Requestor's Name, Employee ID #*

Title:

Phone #

(NOTE: Certificate of officer authorized by law to take acknowledgements is not essential to the validity of Notice of Federal Tax Lien, Rev. Rul. 71-466, 1971-2 C.B. 409)

Part 1 – Recording Office Form 668(Y) (Rev. 2-2004) Cat. No. 60025X *Image credit to the* Internal Revenue Service

Pictured above is a template for a federal tax lien, which is a claim of ownership against one's property for failure to pay a tax debt (in this case a foreign debt on behalf of a "Treaty Partner," which includes foreign entities such as Canada, Denmark, France, Netherlands, and Sweden).

A federal tax lien is an example of an adversarial claim of ownership which complicates ownership and transfer of titles due to potential foreign legal complications. However, if one can find its

corresponding release, which is a record which proves the adversarial claim has been “released,” the adversarial claim is neutralized, and one’s claim of ownership is no longer challenged.

With the help of optical character recognition (OCR), challenges such as this one can be solved, as the following section discusses.

Building an End-to-End Pipeline

Imagine a bank that wants an end-to-end pipeline where they can input a raw scanned image of a deed and obtain useful information, which proves that the bank and its customers have strong legal claims to their rights and properties that can survive legal challenge.

The first step in the pipeline is to filter out bad documents. Many documents are of poor image quality due to being old, from fire or water damage, or simply a bad scan operation. Documents that are too dirty to be processed by a computer must be filtered out and interpreted by a trained human instead. A combination of image quality metric and optical character recognition (OCR) quality metric should be sufficient; we suggest using image noise, and how many OCR errors occur, during a preliminary OCR as metrics.

The second step is to digitize the document into text with optical character recognition, which is easier said than done. There are a number of preprocessing steps to obtain the best optical character recognition results. Documents should be transformed with intensive de-noising, de-rotation, gamma correction, and text correction. However, text correction should not be too aggressive – it is better to leave an error uncorrected than to, for example, “correct” a named entity such as a foreign investor with a non-English name. Rather, one should be strategic on which words should be corrected and which shouldn’t, and this is heuristic work that requires domain expert knowledge.

The next step is to classify these documents. There are a variety of deeds such as quitclaim deeds, warranty deeds, and trust deeds, coming from various counties or districts. The ability to separate them not only improves accuracy further down the pipeline but can also be used for error checking manually submitted documents with incorrect data. One should leverage natural language and computer vision features, but one can also leverage unique visual features such as stamps and barcodes since different counties and districts use different stamps and place barcodes in different areas on the document.

Finally, we wish to extract useful information from the document such as “who is the owner,” “who is the borrower,” “how much is owed,” and “who is the insurer” because they help document the chain of ownership of a property. Conveniently, all these answers can be found right on the deed, so this becomes a question-answering task – locating where the answer exists on the document based on context. There are many ways to approach a question-answering task, but one way is a Bidirectional Encoder Representations from Transformers (BERT) model pre-trained on a real estate legal corpus and tuned on this Q&A task.

Conclusion

We have discussed a plausible end-to-end pipeline for solving a problem extracting useful information from a massive corpus of undigitized and unlabeled documents, to provide proof of property ownership for countless individuals serving as customers for a bank and the bank’s customers. There does not exist a centralized database for American deeds of title ownership. But this challenge is also an opportunity to leverage data science expertise to provide value for our

clients. Here is where Centific can help. At Centific, we strive to meet the unique data science challenges and opportunities within broad environments not only in real estate, but also retail, finance, and other industries. Contact us to learn how we can help you.

-
-
-
-
-