Bankruptcy Map: A System for Searching and Analyzing U.S. Bankruptcy Cases at Scale

Euirim Choi
euirim@cs.stanford.edu
Stanford University
Stanford, California

Gillian Brassil
gbrassil@stanford.edu
Stanford University
Stanford, California

Katie Keller
ktkeller@stanford.edu
Stanford University
Stanford, California

Jessica Ouyang
ouyangj@stanford.edu
Stanford University
Stanford, California

Kate Wang
katewang@stanford.edu
Stanford University
Stanford, California

ABSTRACT

With many involving tens if not hundreds of thousands of pages of legal documents, bankruptcy cases can be overwhelming for many reporters to tackle. After informally interviewing industry experts specializing in investigating bankruptcies, we developed the Bankruptcy Map, a system that can help journalists dive deeper in a bankruptcy case of interest and spot trends across bankruptcies. With the help of machine learning models and process parallelization, the Bankruptcy Map can efficiently and intelligently extract text from often poorly formatted and irregular legal documents, perform named entity recognition to identify key actors in a particular case, and enumerate the creditors associated with a particular case via parsing paper forms. All of the information processed by the Bankruptcy Map can be accessed by a journalist through an intuitive online interface built with the latest web technologies. When evaluated on a subset of chapter 11 bankruptcies from a popular jurisdiction in the United States, we found that our system was effective at all its tasks and we were able to use the tool to help identify a newsworthy trend.

KEYWORDS

bankruptcy, creditors, clustering, journalism, law, named entity recognition, trendspotting, human computer interaction

ACM Reference Format:
Euirim Choi, Gillian Brassil, Katie Keller, Jessica Ouyang, and Kate Wang. 2020. Bankruptcy Map: A System for Searching and Analyzing U.S. Bankruptcy Cases at Scale. In *Computation + Journalism Symposium 2020, March 20–21, 2020, Boston, MA.* 5 pages.

C+J '20, March 20–21, 2020, Boston, MA

1 INTRODUCTION

Every year, hundreds of thousands of bankruptcies are filed in the United States [5, 8]. While many are individual bankruptcies filed by people unable to payback their loans, a little over 20,000 cases are filed by businesses [5].

When a business, especially those with large market capitalizations, files for bankruptcy, the effects are often disruptive to the wider economy: employees are laid off, pension funds can become insolvent, and markets can experience turbulence. Bankruptcies are therefore often newsworthy, and a cadre of journalists across the country have covered the space regularly in recent years [9, 10].

Bankruptcy reporting, however, is not without its challenges. Corporate bankruptcies, especially those involving multinationals with a plethora of subsidiaries, can involve hundreds of thousands of pages of documents, making it difficult for reporters to find trends in this deluge of data [10].

Some work has been done to make sense out of these massive corpora, particularly in information science, where practioners have built pipelines to intelligently search and parse dense text [18]. In the journalism space, reporters and developers at the International Consortium of Investigative Journalists used named entity recognition and graph databases to pinpoint newsworthy entity relationships in multiple massive dumps of legal documents regarding offshore tax havens [6].

But no specialized open-source solution exists to tackle the analysis of bankruptcy cases at scale. In this paper, we therefore introduce a novel user-friendly tool known as the Bankruptcy Map. The tool can, given a potentially large list of chapter 11 bankruptcy cases, even with a modest budget, parse important information from case metadata like case duration, extract text from associated legal documents, and enumerate important named entities as well as creditors from this text. It can also cluster cases to allow for users to find similar cases to the case of interest.

When tested on a dataset of bankruptcy cases from the United States Bankruptcy Court for the Southern District of New York (NYSB) provided by a non-profit, our tool showed satisfactory performance on all its major functions when reviewed manually on a subset of the data. We were also able to identify a novel trend that was newsworthy with assistance from the tool: chapter 11 bankruptcy cases in the NYSB jurisdiction have been getting shorter in recent years.

2 METHODOLOGY

Interviewing Industry Experts

We first interviewed a bankruptcy reporter at *The Wall Street Journal* and engineers working on facilitating the analysis of bankruptcies at the Dow Jones & Company to understand the nuances of bankruptcies and narrow down the requirements for our tool. We concluded from these conversations that bankruptcy reporters would find a tool most useful if it could allow for the finding of high-level trends in a dataset, as well as the ability to see more granular information like the creditors involved in a particular case. These conversations helped us design the rest of the pipeline.

Data Acquisition

In the United States, federal bankrupcty court documents are filed in the electronic database known as *PACER*, which is managed by the U.S. Courts [11]. While PACER’s records are comprehensive, the service charges up to $3 to download each document. Consequently, to download tens of thousands of documents, as we would need to do, would have been unaffordable with our budget (and to that of most newsrooms).

After an exhaustive review of alternative bankruptcy case database services, we arrived at using bankruptcy case data from RECAP, a database managed by the non-profit Free Law Project that contains case dockets and corresponding documents automatically collected by PACER users who voluntarily install a browser extension. RECAP offers its data in bulk to researchers at no cost.

RECAP provided us the dockets of over two million cases from every federal bankruptcy court in the U.S. in JSON format. In the data, each case had an attached docket composed of entries that enumerated their associated court documents, some of which were available to download for free.

Case Filtering

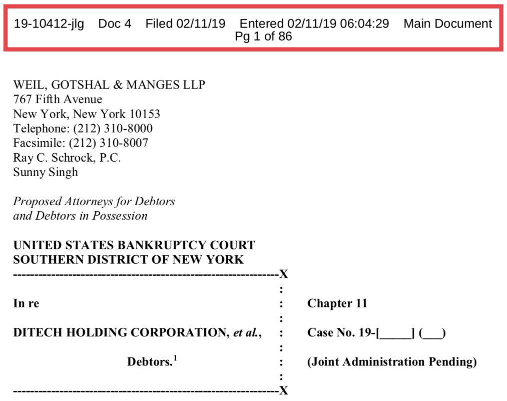
Upon review of the dataset, we found that different jurisdictions often had widely different case document formats, particularly for legal forms. Because intelligently and accurately parsing data from these forms was critical in our project, we opted to restrict our dataset to cases filed in the United States Bankruptcy Court for the Southern District of New York (NYSB). We chose the Southern District because, with jurisdiction over Manhattan in New York City, this court handled some of the largest and most prominent bankruptcy cases in the country [7].

After filtering out the non-NYSB cases, we next excluded non-chapter 11 bankruptcies, based on feedback from two bankruptcy reporters who said that chapter 11 bankruptcies, which more often involve corporate entities, are much more interesting than the others. We used simple heuristics using case metadata to do the filtering, as the chapter number of each case was not provided by RECAP.

After this filtering process, we ended up having 1,968 bankruptcy cases, with over 500,000 associated documents. The earliest case in our dataset dated back to 1993, whereas the most recent was first filed in 2019.

Document Text Extraction

Around 60,000 of these documents were freely accessible in PDF format; we extracted text from them through a three-step process: **(1)** converting the documents into JPEGs, **(2)** cropping the resulting images vertically to eliminate the standardized legal headers (shown in Figure 1) and page numbers present on nearly every document in the dataset, and **(3)** running these images through Tesseract, Google’s optical character recognition (OCR) engine [17]. We chose to use the long short-term memory (LSTM) neural network engine available since Tesseract version 4, due to its superior performance [12, 13].



19-10412-jlg Doc 4 Filed 02/11/19 Entered 02/11/19 06:04:29 Main Document
Pg 1 of 86

WEIL, GOTSHAL & MANGES LLP
767 Fifth Avenue
New York, New York 10153
Telephone: (212) 310-8000
Facsimile: (212) 310-8007
Ray C. Schrock, P.C.
Sunny Singh

Proposed Attorneys for Debtors
and Debtors in Possession

UNITED STATES BANKRUPTCY COURT
SOUTHERN DISTRICT OF NEW YORK

In re	:	**Chapter 11**
DITECH HOLDING CORPORATION, ***et al.*****,**	:	**Case No. 19-[_____] (___)**
Debtors.[1]	:	**(Joint Administration Pending)**

Figure 1: Legal documents on PACER almost always have a standard header outlined in the red box. Not cropping out headers can adversely affect the quality of the OCR output.

As many bankruptcy documents are hundreds, and sometimes thousands, of pages long, extracting text from all the documents in our dataset within a narrow timeframe with minimal resources was non-trivial, especially using the slower LSTM engine in Tesseract. We ended up parallelizing our text extraction pipeline and running the entire process on an Amazon Web Services (AWS) EC2 instance with 16 cores, each supporting four threads.

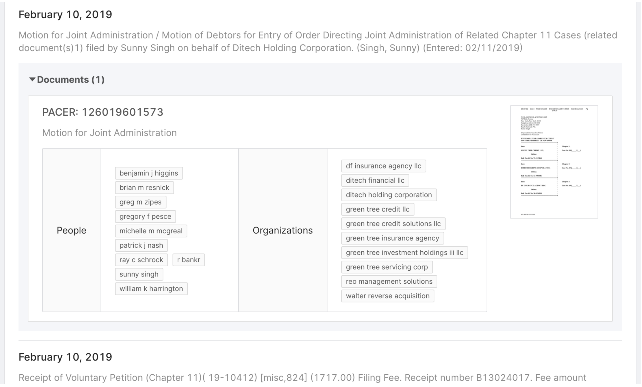
Besides performance concerns, we ran into other minor obstacles that were eventually resolved, including, but not limited to, the fact that many legal documents used proprietary Microsoft fonts that were unrecognizable in our Linux environment, degrading the quality of text extraction.

Some documents were unreadable by Tesseract, often due to being unusually oriented (i.e. backwards, upside down, etc.). To resolve this problem, we discarded documents that had text that did not have at least 65% of known English words.

Named Entity Recognition

We then ran named entity recognition on each document's extracted text using a spaCy convolutional neural network that uses a bloom embedding strategy with convolutional layers [14]. The network was pre-trained on the large OntoNotes dataset, with GloVe vectors used for feature creation trained on Common Crawl data [3, 16].

The model recognized named entities and their types. We collected entities identified as being either people or organizations and discarded the rest. This was because we found through manual review that the model struggled to accurately identify the types of other named entities.



February 10, 2019

Motion for Joint Administration / Motion of Debtors for Entry of Order Directing Joint Administration of Related Chapter 11 Cases (related document(s)1) filed by Sunny Singh on behalf of Ditech Holding Corporation. (Singh, Sunny) (Entered: 02/11/2019)

▾Documents (1)

PACER: 126019601573

Motion for Joint Administration

People	benjamin j higgins brian m resnick greg m zipes gregory f pesce michelle m mcgreal patrick j nash ray c schrock r bankr sunny singh william k harrington	Organizations	df insurance agency llc ditech financial llc ditech holding corporation green tree credit llc green tree credit solutions llc green tree insurance agency green tree investment holdings iii llc green tree servicing corp reo management solutions walter reverse acquisition

February 10, 2019

Receipt of Voluntary Petition (Chapter 11)(19-10412) [misc,824] (1717.00) Filing Fee. Receipt number B13024017. Fee amount

Figure 2: An example of named entity extraction on a legal document displayed via the online interface.

Initially, we tried using Blackstone, a third-party spaCy model that supposedly specializes in legal texts, to identify entity types that were specific to the legal world [15]. We found, however, that the model struggled on American bankrupty case documents, likely because it was trained on legal documents from England and Wales.

Case Creditor Extraction

In parallel to our Tesseract-based document text extraction pipeline, we extracted the creditors associated with each bankruptcy, a piece of information that Corrigan from the *Journal* and staff from Dow Jones identified as being critical to bankruptcy reporting.

While creditor information could not be determined directly from the RECAP data, we took advantage of a feature common to chapter 11 bankruptcy cases: the Form 204. Filed at the beginning of each case's docket, this official form requires the party seeking bankruptcy to identify its major creditors in table format, along with the amounts they are owed [1].

After isolating the documents that contained Form 204s using keyword detection, we passed over 20,000 pages of these files through AWS' Textract, which uses computer vision to parse table data coherently [4]. This process was the most expensive step in the race to build the tool; it cost around $340. We discarded the amounts that the extracted creditors were owed as the data was too imprecise upon manual review of a subset of the cases.

Online Interface

After filtering and processing the data, we uploaded the data to a PostgreSQL relational database. We built a backend server using Django, a web framework written in Python, that used Elasticsearch to dramatically speed up case searching. Both the database and backend servers were hosted on AWS.

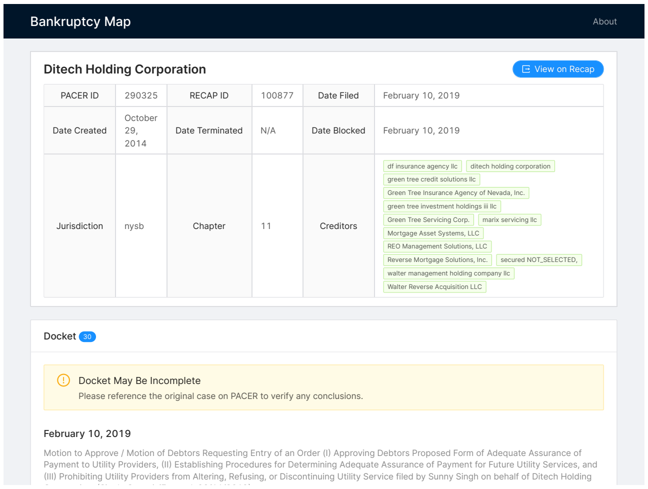
The backend communicates with our frontend written in JavaScript and React to display data from the database to users through an intuitive interface. Users are able to search for cases in the database and click on them to see details extracted from the case documents and metadata, as seen in Figure 3. They are also able to peruse each case's docket and examine the original PDFs of the associated documents as well as their extracted named entities. By clicking on one of these named entities, users are able to see all the cases that are associated with that entity.

Case Clustering

By storing cases and named entities extracted from their documents in a relational database, we were able to dynamically calculate, when given a case, cases that are semantically similar. We did this by getting cases that have the most similar number of entities to a given case and displaying the top three.

3 RESULTS

Getting empirical results out of our technology with a limited budget was difficult due to the scale of the data being processed in conjunction with the lack of a ground truth for our tasks. As a result, we decided to evaluate our tool's accuracy via manual review of a random sample of our dataset.



Ditech Holding Corporation

PACER ID	290325	RECAP ID	100877	Date Filed	February 10, 2019
Date Created	October 29, 2014	Date Terminated	N/A	Date Blocked	February 10, 2019
Jurisdiction	nysb	Chapter	11	Creditors	df insurance agency llc ditech holding corporation green tree credit solutions llc Green Tree Insurance Agency of Nevada, Inc. green tree investment holdings iii llc Green Tree Servicing Corp. marix servicing llc Mortgage Asset Systems, LLC REO Management Solutions, LLC Reverse Mortgage Solutions, Inc. secured NOT_SELECTED, walter management holding company llc Walter Reverse Acquisition LLC

Docket 30

Docket May Be Incomplete
Please reference the original case on PACER to verify any conclusions.

February 10, 2019

Motion to Approve / Motion of Debtors Requesting Entry of an Order (I) Approving Debtors Proposed Form of Adequate Assurance of Payment to Utility Providers, (II) Establishing Procedures for Determining Adequate Assurance of Payment for Future Utility Services, and (III) Prohibiting Utility Providers from Altering, Refusing, or Discontinuing Utility Service filed by Sunny Singh on behalf of Ditech Holding

Figure 3: An example page on the online interface that shows the details of a particular bankruptcy case, including associated creditors and the docket.

To evaluate case filtering, we looked at 25 random cases that our model classified as chapter 11 in our dataset and found that 23 were correctly labelled as such. As a proxy for recall, we measured the proportion of total cases that were labelled as chapter 11. We found that roughly 10% of NYSB cases were marked as being chapter 11, which is higher than the national average of a little less than one percent. We think this is explained by the fact that the NYSB hears more bankruptcies than nearly every other jurisdiction in the United States [8]. It may also be explained by the fact that RECAP acquires cases through users, often journalists, using its browser extension to look at newsworthy cases, which may lead to an overrepresentation of corporate bankruptcies.

We extracted text successfuly from around 70,000 documents, and, upon manually reviewing twenty random documents, we found that 14 had extracted text that we deemed acceptable. We were able to extract named entities from over 36,000 documents and, given a random sample of 10 of these documents, we determined that our program correctly interpreted each entity 92% of the time; it found all the important entities in a document 85% of the time. Using these entities to cluster cases led to a correct top-3 similar case identification accuracy of 70% in a random sample of 10 cases. Finally, we managed to extract creditors from 593 cases. Out of 20 of these cases that we reviewed by hand after analysis, 15 had at least one false positive or false negative for creditors; still, 17 of them had a majority of their creditors present.

4 EVALUATION

We used the Bankruptcy Map to find a newsworthy trend in our dataset in order to demonstrate its potential utility to journalists. A *Wall Street Journal* reporter is planning on using our finding in future reporting.

	log HR	log HR SE	HR	t	P >|t|
year	0.1870	0.0191	1.2056	9.8116	0.0000

Table 1: Proportional hazards regression results showing the relationship between the year a case was filed and its duration.

Specifically, we examined whether cases were getting longer or shorter over time. Normally, examining the relationship between two variables would immediately prompt thoughts of using a common form of linear regression like ordinary least squares. In this problem, however, cases that were filed more recently are more likely to be ongoing simply because they are more recent. Disposing such cases due to missing data would consequently bias any traditional linear regression model.

To account for this phenomenon, we opted to use *proportional hazards regression*, a statistical model that takes into account *censored*, or partially known, data in order to reduce bias [2]. The results of the regression is found in Table 1. The larger than one coefficient of the hazard ratio (HR), indicates that bankruptcy cases in the Southern District of New York have terminated somewhat quicker in recent years. The result is statistically significant.

5 CONCLUSION & FUTURE WORK

With our tool already able to be used to discover a newsworthy trend regarding the decreasing length of chapter 11 bankruptcy cases in recent years, the Bankruptcy Map, a pipeline of technologies packaged to make the process of analyzing large troves of documents efficiently and accurately, has the potential to be a useful news discovery tool to bankruptcy journalists across the country.

While the tool is already useful in trendspotting and clustering, it can be improved with further work. Using named entity recognition models trained on labelled U.S. legal, and specifically bankruptcy, data would likely result in higher-quality named entities being extracted, potentially improving clustering accuracy. Other natural language processing algorithms, such as automated summarization, could be used to make it easier for reporters to process large numbers of documents.

Any of these next steps could make the Bankruptcy Map more useful when analyzing huge corpora; the underlying technologies may also prove to be generalizable to other domains outside the bankruptcy space that feature large troves of highly technical documents.

ACKNOWLEDGMENTS

We gratefully acknowledge Director of the Brown Institute for Media Innovation Maneesh Agrawala, Managing Director of the Stanford Journalism and Democracy Initiative R.B. Brenner, and Google News founder Krishna Bharat for their guidance on this project, and to master's student Kylie Jue and Lorry I. Lokey Visiting Professor in Journalism Serdar Tumgoren for their support. We thank Tom Corrigan from *The Wall Street Journal* and Dylan Roy, Jonathan Duxbury, and Wiktor Szary from Dow Jones and Company for inspiring and advising us. Finally, we thank the Free Law Project for graciously allowing us to use data from the RECAP project free of charge.

REFERENCES

[1] [n.d.]. Official Form 204. https://www.uscourts.gov/sites/default/files/form_b204_0.pdf

[2] 1996. Extending the Cox Model. *Technical Report Series Section of Biostatistics* (11 1996). https://www.mayo.edu/research/documents/biostat-58pdf/DOC-10027288

[3] 2016. spaCy Models Documentation. https://spacy.io/models

[4] 2019. Amazon Textract. https://aws.amazon.com/textract/

[5] 2019. Bankruptcy Filings Continue to Decline. https://www.uscourts.gov/news/2019/04/22/bankruptcy-filings-continue-decline

[6] Uli and Foessmeier. 2016. Uncovering Secrets in the Offshore Leaks Database with Tom Sawyer Perspectives (Part 1) - Neo4j Graph Database Platform. https://neo4j.com/blog/offshore-leaks-database-tom-sawyer-perspectives/

[7] Richard and Maidman. 2012. Retelling the History of the United States District Court for the Southern District of New York. , 16-17 pages. https://www.pbwt.com/content/uploads/2015/07/NYLitigator-Summer2012-History-SDNY-1043294-_2_.pdf

[8] George N. and Panagakis. 2015. Trends in Chapter 11 Filings, Venue and Proposed Reforms | Insights | Skadden, Arps, Slate, Meagher & Flom LLP. https://www.skadden.com/insights/publications/2015/01/trends-in-chapter-11-filings-venue-and-proposed-re

[9] Tom Corrigan. 2018. The Power Players That Dominate Chapter 11 Bankruptcy. *Wall Street Journal* (01 2018). https://www.wsj.com/graphics/bankruptcy-power-players/

[10] Tom Corrigan. 2019. Bankruptcy Data Shows Concentrated Industry. https://www.wsj.com/articles/bankruptcy-data-shows-concentrated-industry-11558713807

[11] United States Courts. 2019. Public Access to Court Electronic Records. https://www.pacer.gov

[12] Google. 2017. 4.0 Accuracy and Performance. https://github.com/tesseract-ocr/tesseract/wiki/4.0-Accuracy-and-Performance

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9 (11 1997), 1735–1780.

[14] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[15] ICLRandD. 2019. Blackstone. https://github.com/ICLRandD/Blackstone

[16] Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. OntoNotes: A Unified Relational Semantic Representation. *International Journal of Semantic Computing* 01 (12 2007), 405–419.

[17] R. Smith. 2007. An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* 2 (09 2007). https://ieeexplore.ieee.org/abstract/document/4376991/citations#citations

[18] Gian Piero Zarri. 1983. RESEDA, an Information Retrieval system using artificial intelligence and knowledge representation techniques. *ACM SIGIR Forum* 17 (06 1983), 189.

A ONLINE RESOURCES

The code for our tool is found at https://github.com/euirim/bankruptcy-db. Find our analysis code at https://github.com/euirim/bankruptcy-map. Visit https://www.courtlistener.com/recap to learn more about RECAP.