On the Books: Jim Crow and Algorithms of Resistance

White paper

Introduction

On the Books: Jim Crow and Algorithms of Resistance is a collections as data and machine learning project of the University of North Carolina at Chapel Hill Libraries. The first phase of the project was funded through *Collections as Data Part to Whole*, made possible by The Andrew W. Mellon Foundation. The project created text corpora of North Carolina session laws and utilized machine learning techniques to discover Jim Crow laws passed in the period between Reconstruction and the Civil Rights Movement (1865-1967). Products from the project include two text corpora: one of all session laws passed during the period of study, and another of all laws during the period that were identified as those likely to be Jim Crow laws. This white paper describes the methods and workflows used to create the corpora and the text analysis techniques applied to identify the Jim Crow laws for the first phase of the project (ending August 31, 2020). All programming work was done using Python. Documented scripts and examples can be found on the project's GitHub page (https://github.com/UNC-Libraries-data/OnTheBooks).

Corpus creation

Collecting Images for OCR

The collection used for this project was digitized between 2009-2011 under the IMLS grant *Ensuring Democracy through Digital Access*, a partnership between East Carolina University, the State Library of North Carolina, and the University Libraries at the University of North Carolina at Chapel Hill. The project digitized state documents and made them accessible through the Internet Archive. *On the Books* expanded on this digitization work by using optical character recognition (OCR) on the images to create a text corpus and by adding metadata. The first step in creating the corpus was to generate a list of all volumes from the period of interest. The corpus includes all volumes from 1866/67-1967. See Appendix 1 for a complete list of volumes included.

As the UNC University Libraries digitized the volumes for *Ensuring Democracy through Digital Access*, UNC bibliographic record IDs were included in the Internet Archive metadata. The Internet Archive's advanced search interface [https://archive.org/advancedsearch.php] allows for the export of metadata for search results in a .csv format. The bibliographic record IDs were used in an advanced search to return Internet Archive metadata for the volumes.

Here is a truncated example of the advanced search used to return metadata for the volumes:

```
scanningcenter:chapelhill AND mediatype:texts AND
unc_bib_record_id:( b4138587 OR b3141545…)
```

The search produced a csv file (search.csv), containing the Internet Archive unique identifiers, titles, and dates for each volume to be included in the corpus. A Python script was written to download JP2 images and jpegs of all pages in all volumes, based on the unique identifiers. The download script references the search.csv file, which provides the list of unique IDs of the volumes needed.

Identifying images to OCR

The print volumes were digitized in their entirety. Images include blank pages, indexes, title pages, etc., which did not need to be included in the corpus. To identify which pages needed to be put through OCR, jpeg images were reviewed, and those that did not need to be put through OCR were deleted. This labor-intensive process was made more efficient by using jpeg images rather than the JP2s. Image previews were used to identify those that could be deleted. Lists of the remaining files were created and augmented with metadata.

Gathering Metadata

Many of the volumes contained laws from multiple sessions of the general assembly and different types of laws. The following metadata was gathered manually, for each image to be put through OCR:

- filename (created by listing jpegs remaining from previous step)
- page number (page number printed on the original pages)
- law type (public, private, public-local, or session laws)
- transcription of the title of the law type section (e.g. Private Laws of the State of North Carolina, Session 1891)

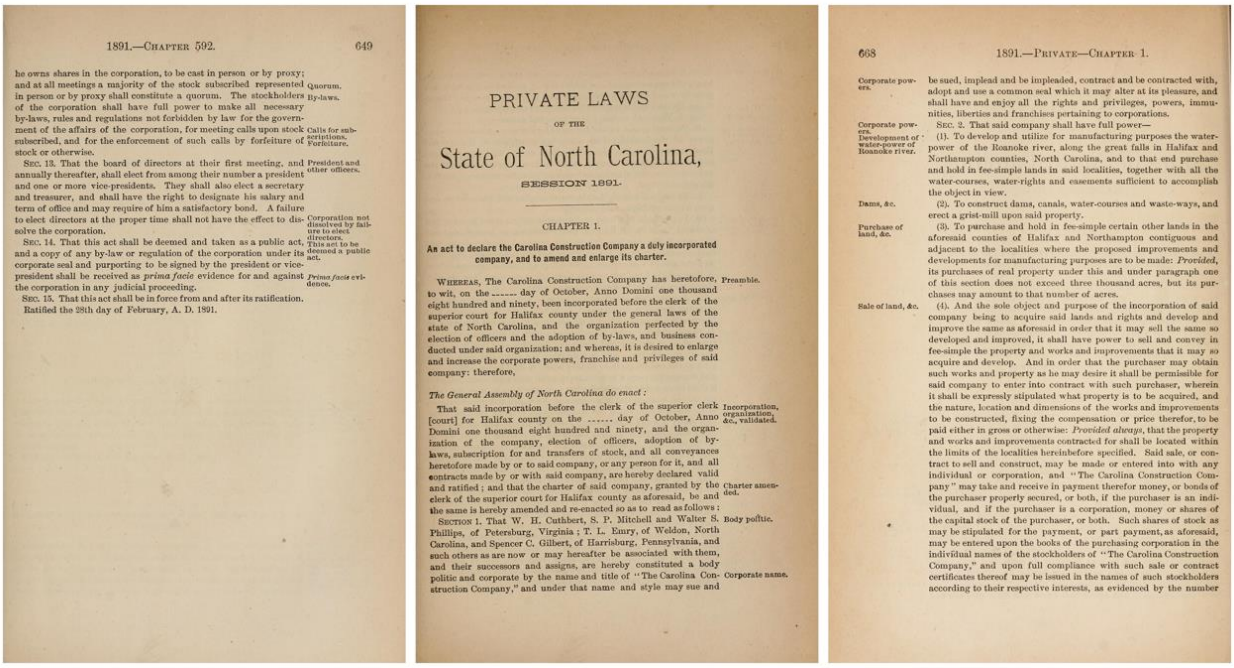
A Python script was used to scrape metadata from the "_scandata.xml" files downloaded from the Internet Archive. The following metadata was scraped:

- leafNum (pdf page number)
- handSide (right or left)

These metadata were used in subsequent processing and/or are included in the XML metadata describing the corpus.

Preparing images for OCR

As shown in the below images, the text on the images are skewed, and the pages contain marginalia and headers. Image quality (color, contrast, etc.) varied across the corpus.



1891.—CHAPTER 592. 649

he owns shares in the corporation, to be cast in person or by proxy; and at all meetings a majority of the stock subscribed represented in person or by proxy shall constitute a quorum. The stockholders of the corporation shall have full power to make all necessary by-laws, rules and regulations not forbidden by law for the government of the affairs of the corporation, for meeting calls upon stock subscribed, and for the enforcement of such calls by forfeiture of stock or otherwise.

Quorum. By-laws.

Calls for subscriptions. Forfeiture.

SEC. 13. That the board of directors at their first meeting, and annually thereafter, shall elect from among their number a president and one or more vice-presidents. They shall also elect a secretary and treasurer, and shall have the right to designate his salary and term of office and may require of him a satisfactory bond. A failure to elect directors at the proper time shall not have the effect to dissolve the corporation.

President and other officers.

Corporation not dissolved by failure to elect directors.

SEC. 14. That this act shall be deemed and taken as a public act, and a copy of any by-law or regulation of the corporation under its corporate seal and purporting to be signed by the president or vice-president shall be received as *prima facie* evidence for and against the corporation in any judicial proceeding.

This act to be deemed a public act.

Prima facie evidence.

SEC. 15. That this act shall be in force from and after its ratification.

Ratified the 28th day of February, A. D. 1891.

PRIVATE LAWS

OF THE

State of North Carolina,

SESSION 1891.

CHAPTER 1.

An act to declare the Carolina Construction Company a duly incorporated company, and to amend and enlarge its charter.

Preamble.

WHEREAS, The Carolina Construction Company has heretofore, to wit, on the ..... day of October, Anno Domini one thousand eight hundred and ninety, been incorporated before the clerk of the superior court for Halifax county under the general laws of the state of North Carolina, and the organization perfected by the election of officers and the adoption of by-laws, and business conducted under said organization; and whereas, it is desired to enlarge and increase the corporate powers, franchise and privileges of said company: therefore,

The General Assembly of North Carolina do enact:

Incorporation, organization, &c., validated.

That said incorporation before the clerk of the superior clerk [court] for Halifax county on the ..... day of October, Anno Domini one thousand eight hundred and ninety, and the organization of the company, election of officers, adoption of by-laws, subscription for and transfers of stock, and all conveyances heretofore made by or to said company, or any person for it, and all contracts made by or with said company, are hereby declared valid and ratified; and that the charter of said company, granted by the clerk of the superior court for Halifax county as aforesaid, be and the same is hereby amended and re-enacted so as to read as follows:

Charter amended.

Body politic.

SECTION 1. That W. H. Cuthbert, S. P. Mitchell and Walter S. Phillips, of Petersburg, Virginia; T. L. Emry, of Weldon, North Carolina, and Spencer C. Gilbert, of Harrisburg, Pennsylvania, and such others as are now or may hereafter be associated with them, and their successors and assigns, are hereby constituted a body politic and corporate by the name and title of "The Carolina Construction Company," and under that name and style may sue and

Corporate name.

668 1891.—PRIVATE—CHAPTER 1.

Corporate powers.

be sued, implead and be impleaded, contract and be contracted with, adopt and use a common seal which it may alter at its pleasure, and shall have and enjoy all the rights and privileges, powers, immunities, liberties and franchises pertaining to corporations.

Corporate powers. Development of water-power of Roanoke river.

SEC. 2. That said company shall have full power—

(1). To develop and utilize for manufacturing purposes the water-power of the Roanoke river, along the great falls in Halifax and Northampton counties, North Carolina, and to that end purchase and hold in fee-simple lands in said localities, together with all the water-courses, water-rights and easements sufficient to accomplish the object in view.

Dams, &c.

(2). To construct dams, canals, water-courses and waste-ways, and erect a grist-mill upon said property.

Purchase of land, &c.

(3). To purchase and hold in fee-simple certain other lands in the aforesaid counties of Halifax and Northampton contiguous and adjacent to the localities where the proposed improvements and developments for manufacturing purposes are to be made: *Provided*, its purchases of real property under this and under paragraph one of this section does not exceed three thousand acres, but its purchases may amount to that number of acres.

Sale of land, &c.

(4). And the sole object and purpose of the incorporation of said company being to acquire said lands and rights and develop and improve the same as aforesaid in order that it may sell the same so developed and improved, it shall have power to sell and convey in fee-simple the property and works and improvements that it may so acquire and develop. And in order that the purchaser may obtain such works and property as he may desire it shall be permissible for said company to enter into contract with such purchaser, wherein it shall be expressly stipulated what property is to be acquired, and the nature, location and dimensions of the works and improvements to be constructed, fixing the compensation or price therefor, to be paid either in gross or otherwise: *Provided always*, that the property and works and improvements contracted for shall be located within the limits of the localities hereinbefore specified. Said sale, or contract to sell and construct, may be made or entered into with any individual or corporation, and "The Carolina Construction Company" may take and receive in payment therefor money, or bonds of the purchaser properly secured, or both, if the purchaser is an individual, and if the purchaser is a corporation, money or shares of the capital stock of the purchaser, or both. Such shares of stock as may be stipulated for the payment, or part payment, as aforesaid, may be entered upon the books of the purchasing corporation in the individual names of the stockholders of "The Carolina Construction Company," and upon full compliance with such sale or contract certificates thereof may be issued in the names of such stockholders according to their respective interests, as evidenced by the number

To obtain the highest quality OCR, JP2 images were edited to straighten text and image parameters such as color and contrast were adjusted to optimize OCR. Headers, footers, and marginalia were removed from images.

For each image, the angle of rotation from the horizontal (skew) was identified. Each image was rotated as needed to ensure horizontal text alignment.

Marginalia, which is text that serves as a finding aid, was printed in the corpus volumes prior to 1951. The marginalia are not part of the laws and needed to be left out of the OCR process. The marginalia are not in the same location on each page and are quite close to the text of the laws. Page headers and footers (generally consisting of the chapter, page number, year, and law type) also needed to be excluded from OCR. Tesseract OCR, the software used for OCR, performs best when the text is not too close to the edge of the page ("Tessdoc," n.d., Tesseract OCR Documentation, last updated August 11, 2020). Prior to OCR, each image was trimmed to exclude anything other than the main text body, and a 200-pixel border that matched the image background color was added. For each page, coordinates of areas to be put through OCR and the median color of the image was recorded for use as input in the OCR script.

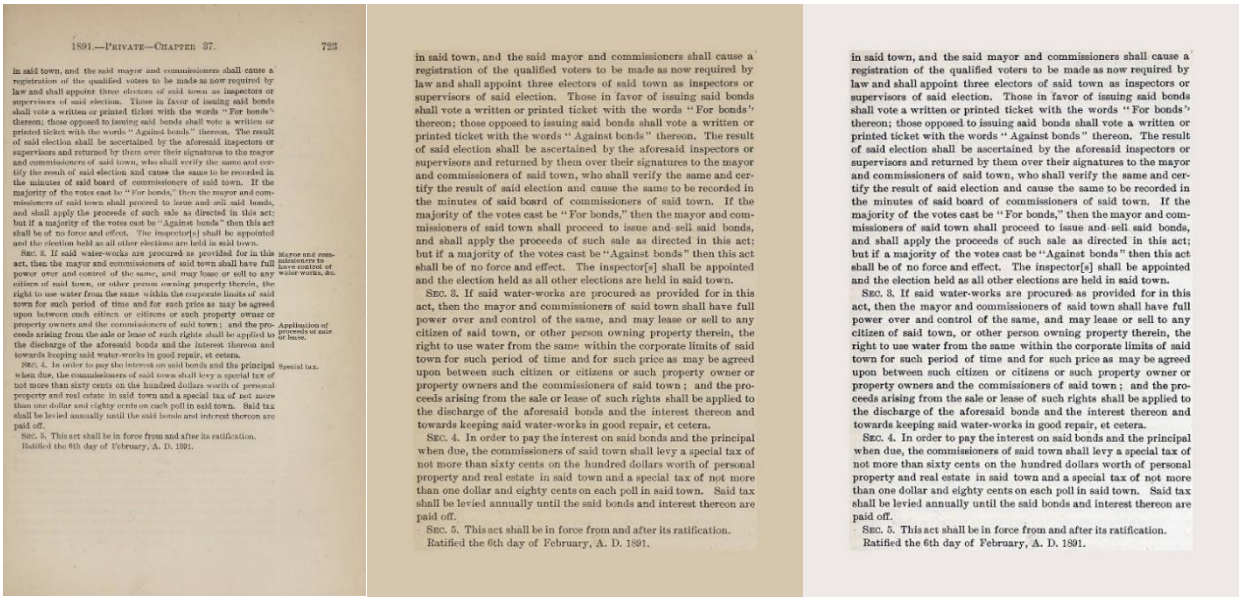
Previously gathered metadata were used in the marginalia and header determination process. The "handSide" values (right or left) were used to determine the side of the image containing marginalia. The first page of each law type section (public laws, private laws, etc.) contains valuable metadata that needed to be included in the OCR. Those pages were identified using the law type metadata. Page images with a different law type from the page immediately before were designated as "start" pages. Headers were retained for the start pages.

Images were adjusted using Python Imaging Library, Pillow 7.0.0 to improve OCR results. To determine image adjustments that would optimize OCR for each volume, a sample of 10 images from each volume was tested. Based on graphic design experience, our team chose the following image adjustment

parameters to test (in order): color, autocontrast, blur, sharpen, smooth, and xsmooth (the xsmooth parameter was only used on images for which smooth was found to improve readability). Each sample image was put through OCR to determine baseline OCR quality. OCR output was tokenized with the Natural Language Toolkit (NLTK) and compared to the SpellChecker dictionary to determine the number of unknown words. North Carolina geographic names from geonames.org were added to improve the results ("NC.zip," n.d., GeoNames, accessed August 25, 2020). For each image, the following OCR quality indicators were recorded: number of tokens, number of unknown words, readability score, and list of unknown words. Readability score was calculated by dividing the total number of tokens by the number of unknown words.

The images were then adjusted iteratively through multiple levels of a single parameter. OCR quality results were recorded for each iteration, and the results were compared for each adjustment. The image adjustment that produced the highest readability score was deemed the optimal adjustment for that parameter. All sample images for the volume were adjusted before the next parameter was tested to account for the possibility that adjustments for one parameter may affect the adjustment needed for the next parameter. Optimal adjustment values were recorded for each image. Prior to OCR, images were adjusted according to the recorded parameter specifications. Ultimately, Tesseract OCR performed quite well on the original images. Image adjustments improved readability very little (0.1% improvement).

The following images demonstrate the image pre-processing steps. The image on the left shows the original image: skewed, with a header, marginalia, and no image improvements. The image in the middle depicts an image that has been cropped to the specifications laid out in the marginalia removal process, with a border matching the background color of the image added around the text body. The image on the right depicts the image that was used for OCR, with image adjustments applied. In this example, the image was adjusted by changing the color value from 1 to .75 and changing autocontrast from 0 to 4.



1891.—PRIVATE—CHAPTER 37. 723

in said town, and the said mayor and commissioners shall cause a registration of the qualified voters to be made as now required by law and shall appoint three electors of said town as inspectors or supervisors of said election. Those in favor of issuing said bonds shall vote a written or printed ticket with the words "For bonds" thereon; those opposed to issuing said bonds shall vote a written or printed ticket with the words "Against bonds" thereon. The result of said election shall be ascertained by the aforesaid inspectors or supervisors and returned by them over their signatures to the mayor and commissioners of said town, who shall verify the same and certify the result of said election and cause the same to be recorded in the minutes of said board of commissioners of said town. If the majority of the votes cast be "For bonds," then the mayor and commissioners of said town shall proceed to issue and sell said bonds, and shall apply the proceeds of such sale as directed in this act; but if a majority of the votes cast be "Against bonds" then this act shall be of no force and effect. The inspector[s] shall be appointed and the election held as all other elections are held in said town.

Mayor and commissioners to have control of water-works, &c.

SEC. 3. If said water-works are procured as provided for in this act, then the mayor and commissioners of said town shall have full power over and control of the same, and may lease or sell to any citizen of said town, or other person owning property therein, the right to use water from the same within the corporate limits of said town for such period of time and for such price as may be agreed upon between such citizen or citizens or such property owner or property owners and the commissioners of said town; and the proceeds arising from the sale or lease of such rights shall be applied to the discharge of the aforesaid bonds and the interest thereon and towards keeping said water-works in good repair, et cetera.

Application of proceeds of sale or lease.

SEC. 4. In order to pay the interest on said bonds and the principal when due, the commissioners of said town shall levy a special tax of not more than sixty cents on the hundred dollars worth of personal property and real estate in said town and a special tax of not more than one dollar and eighty cents on each poll in said town. Said tax shall be levied annually until the said bonds and interest thereon are paid off.

Special tax.

SEC. 5. This act shall be in force from and after its ratification.

Ratified the 6th day of February, A. D. 1891.

Optical Character Recognition

Python-tesseract was used for OCR (pytesseract version v4.1.0.20190314 was used for all volumes except the 1913 extra session, which was processed using v5.0.0-alpha). Python-tesseract is a Python wrapper for Google's Tesseract OCR engine. The OCR script written for this task performs OCR on the entire corpus, volume by volume. The metadata gathered in previous steps were used during the OCR process. This includes metadata generated manually, metadata gathered from the Internet Archive, metadata generated during the marginalia determination process, and the image adjustment specifications.

The images were separated according to law type sections (public laws, private laws, etc.) so OCR output files would be generated for each law type section in each volume.

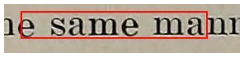
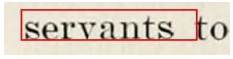
Images were processed using the specifications gathered in the marginalia determination process (rotation, bounding box for the text body to be put through OCR, median background color) and the image adjustment identification process (optimal adjustments for color, autocontrast, blur, sharpen, smooth, and xsmooth). OCR was performed on the prepared images.

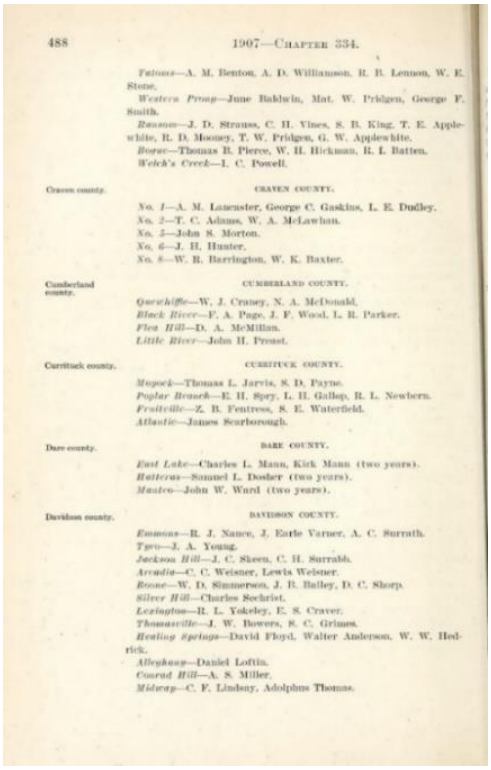
Several outputs were generated from the OCR process, including a text file of the image adjustments made to each volume, a text file of the text recognized by OCR for each law type section, and a word-level .tsv file for each law type section. The .tsv file was used for splitting the volumes into chapters and sections.

OCR Quality Assessment

Two different techniques were used to assess the quality of the OCR. The first technique assessed the page-level OCR quality, and the second assessed the word-level OCR quality.

The quality assessment identified several types of errors:

- Some areas of text were skipped during OCR, resulting in gaps in the text.
- Some areas of text were erroneously excluded due to incorrect marginalia determination.
- Some words were not recognized correctly because they were not delineated correctly by Tesseract:  recognized as "convenient").
- Some words were delineated correctly, but not recognized correctly:  (recognized as "spirituous").
- Pages that show text in tabular format (tables) did not OCR well.



488 1907—CHAPTER 334.

Tatums—A. M. Benton, A. D. Williamson, R. B. Lennon, W. E. Stone.
Western Prong—June Baldwin, Mat. W. Pridgen, George F. Smith.
Bannon—J. D. Strauss, C. H. Vines, S. B. King, T. E. Applewhite, R. D. Mooney, T. W. Pridgen, G. W. Applewhite.
Boyce—Thomas B. Pierce, W. H. Hickman, R. L. Batten.
Welch's Creek—I. C. Powell.

Craven county. CRAVEN COUNTY.

No. 1—A. M. Lancaster, George C. Gaskins, L. E. Dudley.
No. 2—T. C. Adams, W. A. McLawhun.
No. 5—John S. Morton.
No. 6—J. H. Hunter.
No. 8—W. B. Barrington, W. K. Baxter.

Cumberland county. CUMBERLAND COUNTY.

Quewhiffle—W. J. Cranney, N. A. McDonald.
Black River—F. A. Page, J. F. Wood, L. R. Parker.
Flea Hill—D. A. McMillan.
Little River—John H. Proust.

Currituck county. CURRITUCK COUNTY.

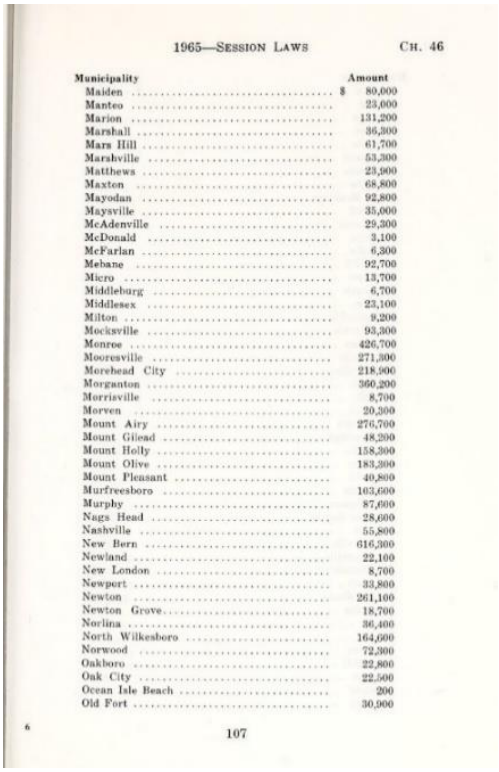
Maple—Thomas L. Jarvis, S. D. Payne.
Poplar Branch—E. H. Spry, L. H. Gallop, R. L. Newbern.
Fruitville—Z. B. Fentress, S. E. Waterfield.
Atlantic—James Scarborough.

Dare county. DARE COUNTY.

East Lake—Charles L. Mann, Kirk Mann (two years).
Hatteras—Samuel L. Dosher (two years).
Manteo—John W. Ward (two years).

Davidson county. DAVIDSON COUNTY.

Emmons—R. J. Nance, J. Earle Varner, A. C. Surratt.
Tyro—J. A. Young.
Jackson Hill—J. C. Sheen, C. H. Surratt.
Arcadia—C. C. Weisner, Lewis Weisner.
Boone—W. D. Simmerson, J. B. Bailey, D. C. Sharp.
Silver Hill—Charles Sechrist.
Lexington—R. L. Yokeley, E. S. Craver.
Thomasville—J. W. Bowers, S. C. Grimes.
Healing Springs—David Floyd, Walter Anderson, W. W. Hedrick.
Alleghany—Daniel Loftin.
Conrad Hill—A. S. Miller.
Midway—C. F. Lindsey, Adolphus Thomas.



1965—SESSION LAWS CH. 46

Municipality	Amount
Maiden	$ 80,000
Manteo	23,000
Marion	131,200
Marshall	36,300
Mars Hill	61,700
Marshville	53,300
Matthews	23,900
Maxton	68,800
Mayodan	92,800
Maysville	35,000
McAdenville	29,300
McDonald	3,100
McFarlan	6,300
Mebane	92,700
Micro	13,700
Middleburg	6,700
Middlesex	23,100
Milton	9,200
Mocksville	93,300
Monroe	426,700
Mooresville	271,300
Morehead City	218,900
Morganton	360,200
Morrisville	8,700
Morven	20,300
Mount Airy	276,700
Mount Gilead	48,200
Mount Holly	158,300
Mount Olive	183,300
Mount Pleasant	40,800
Murfreesboro	103,600
Murphy	87,600
Nags Head	28,600
Nashville	55,800
New Bern	616,300
Newland	22,100
New London	8,700
Newport	33,800
Newton	261,100
Newton Grove	18,700
Norlina	36,400
North Wilkesboro	164,600
Norwood	72,300
Oakboro	22,800
Oak City	22,500
Ocean Isle Beach	200
Old Fort	30,900

6 107

Page-level OCR Quality Assessment.

To assess the pervasiveness of missing text, sections recognized poorly by OCR, and other page-level errors in the corpus, a random sample of 100 pages was selected for close inspection. Workers compared text on the page images with the text from OCR, looking for skipped text sections or poor OCR of large sections. Word-level OCR was assessed separately, so individual words that were incorrectly recognized were not included in the page quality assessment.

Out of the 100 sample pages, two errors were detected. In one page, the first two lines of text at the top of the page were missing from the OCR due to incorrect marginalia determination. The second line of text on the page is a single word, with the rest of the line blank (see image right). Because of the blank space, the first two lines were mis-identified as part of the page header.

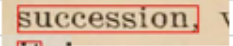
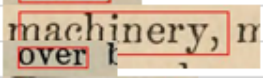
The second error was found in a page that lists municipality names aligned to the left side of the page and dollar amounts aligned to the right side of the page with periods going across the page in between (see image right). As with other pages of text in tabular format, OCR quality was very poor. Much of the information in the corpus that was originally depicted in tabular format will likely be unusable.

The sample size of the page quality assessment is admittedly small. Comparing images with text from OCR word-for-word is very labor intensive. Although it is a small sample, this assessment does provide examples of the types of errors included in the corpus that are consistent with the project team's observations gained through working closely with the corpus. Based on the error rate within the sample, we expect that at least 94% of the pages in the corpus do not have these types of errors.

Word-level OCR Quality Assessment.

To test the quality of word-level OCR, a random sample of 2000 words from the corpus was reviewed for accuracy. Reviewers inspected the words for two types of errors: 1) incorrect word delineation by Tesseract and 2) incorrect word OCR. Tesseract outputs include the bounding box coordinates for each word recognized, and the recognized text. To facilitate assessment, images of the sample words were clipped using Pillow, using the bounding box coordinates returned

from Tesseract plus a small buffer area for readability. The bounding box of each word was outlined in red. The images were added to an Excel spreadsheet, along with the OCR text output (see image below):

	text	wordsampleImageBuffer	is the word deliniation correct (red box)	is the info in the red box OCR'd correctly?
2	or		1	1
3	in		1	1
4	succession,		1	1
5	it		1	1
6	town,		1	1
7	machinery,		1	1
8	over		1	1

Two different reviewers inspected each word in the sample to determine errors. A third person reviewed instances where the two reviewers disagreed. Findings indicate that words were delineated correctly 82.39% of the time, the words were both delineated correctly and recognized correctly 80.43% of the time, and most importantly, the words were recognized correctly 83.76% of the time.

Processing the Text Output

Splitting the text into chapters and sections

The North Carolina session laws are organized by chapter and section. One of the project goals was to use machine learning to identify Jim Crow laws. Since the individual laws are the unit of analysis, the individual laws (or sections) needed to be separated. Much of the time and effort required to build the corpora was spent structuring the data so that individual sections could be analyzed. This was a lengthy process that involved an initial split of the volumes into chapters and sections, numerous iterations of correcting the chapter and section splits, generation of the final split files, and an error assessment.

The relatively standard organization of the law volumes was used to split the chapters and sections. Volumes are organized by law types (private, public, public-local, session laws), each of which is organized by chapters and then sections. Occasional variations from this structure exist but were not captured by the splitting process. Each chapter starts with the word “Chapter,” preceded by blank lines. Each section starts on a new line and begins with “Section” or “Sec.” (in most cases, the word “Section” is spelled out for the first section in each chapter and subsequent sections are abbreviated). Recall that the OCR output included word-level .tsv files for each law type section. The .tsv files were converted to .csv for the splitting process. The splitting process used regex pattern matching to identify the beginning of new chapters and sections. The patterns were identified by investigating volume format and OCR output. Columns for chapter and section were added to the .csv. The splitting script iterates through all rows of each .csv file, and for each row (word), the chapter and section values are added. If the pattern-matching statements identify a new chapter or section, that chapter or section becomes the new respective chapter or section value for all rows beginning with the one in which it was found. This repeats every time a new positive pattern match occurs.

After the initial splits were made, it was clear that there were numerous errors in the chapter and section splits. Examples of errors include regex pattern matching errors, OCR errors, marginalia

determination errors, image scan errors, and errors in the original printed volumes. Although not identified in the OCR quality assessment, it became clear during the splitting process that the numbers of chapters and sections were frequently recognized incorrectly. The numbers 3 and 8 were frequently misrecognized and switched. Four separate efforts to clean up chapter splits were completed—a mix of automated and manual corrections. Chapter splits that were not identified because of missing text (text that was either skipped by the OCR or left out due to errors in the marginalia determination) were corrected by transcribing the missing text. Some volumes could not be split automatically and were split manually. These volumes include those that identify subchapters with Roman numerals, others that have chapter headings in the margins (which were not put through OCR because of marginalia removal), and lastly, some volumes that simply did not split well at all. Due to the extremely time-intensive nature of this work, some split errors remain in the corpus. The corpus of all session laws contains 53,515 chapters and 297,790 sections. Initially, 27,327 chapter/section split errors were identified. 89.7% of the errors were corrected. Time permitting, more cleanup will be done during phase 2 of the project.

Analysis

Unsupervised classification

Latent Dirichlet Allocation (LDA), an unsupervised machine learning technique for topic modelling, was used to generate topics for the laws. This work was done by project team member Rucha Dalwadi and is detailed in her master's paper (Dalwadi 2020). LDA generates topics, returns the probability that the document (or law in this case) belongs to topics, and returns words describing the topics. It is up to the reviewer to interpret the topics.

One topic returned was relevant to the project goal of identifying Jim Crow laws. The words returned for the topic include: township, trustee, school, college, library, high, creek, hill, church, building, park, john, chapel, custodian, white, year, wake, forest, page, university (Dalwadi 2020, 61). This topic was of interest because the word "white" was included in a topic with words describing public places that were often regulated by Jim Crow laws, such as school, library, church, building, park, and university.

Topic modelling is a helpful tool for gaining an understanding of topics represented in a corpus, but it is not a classification technique. To classify laws as either Jim Crow or not Jim Crow, supervised classification was used.

Supervised Classification

Prior to analysis, non-ASCII characters were removed, all characters were converted to lower case, and words split between lines were combined. A list of stop words, very infrequent words and very frequent words were removed from the corpus. Tokenization was performed using NLTK's word_tokenize. Lemmatization was considered but did not meaningfully improve results.

Several different representations of the text were considered. The simplest representation was as a document term matrix, representing each document as raw word frequencies. We also considered a term-frequency inverse document frequency (tf-idf) transformation to reduce the relative weight on words that are important across the corpus (Jones, 1972). Both the raw document term matrix and the tf-idf transformed matrix were tested at full size and after reducing to 500, 100, and 50 components using Principal Components Analysis. Finally, we tested representing the data using the doc2vec

algorithm, using 500 components (Le, 2014). Ultimately, the document term matrix performed better or comparably with other methods and was used in the analysis.

Modeling incorporated metadata about the type of law (public, private, etc.) and the year (coded as the first year in cases of multi-year volumes).

Supervised classification techniques require a training set to train the model how to distinguish between classes. A training set was created, comprised of laws identified by experts as either Jim Crow or not Jim Crow.

To identify Jim Crow laws for the training set, references were compiled from the work of researchers who had identified the laws through close reading. Fifty Jim Crow laws came from Dr. Rev. Pauli Murray's *States Book on Race and Color* (Murray 1951, 329-348), which includes race-based laws from North Carolina session laws and codes. Only session laws that could be matched to the corpus were included. Seventy-seven Jim Crow laws came from legal scholar Dr. Richard Paschal's document entitled "North Carolina Constitutional Provisions and Statutes that Discriminated Based on Race, 1865 to 1920," which will be included as an appendix in a forthcoming publication (Paschal forthcoming).

To identify non-Jim Crow laws for the training set, team members Dr. William Sturkey and Dr. Kimber Thomas assessed a random selection of 1290 laws from the corpus and classified them as either Jim Crow laws (12) or not Jim Crow laws (1278), suggesting a .94% rate of Jim Crow laws during the period of study. However, we believe this rate varies based on the perspective and opinions of the individual assessing laws and their definition of what constitutes a Jim Crow law. In a subset of overlapping evaluations, one of our reviewers agreed with previous assessments of other reviewers roughly 70% of the time. This may reflect differing opinions on the nuanced definition of "Jim Crow," the opaque nature of legal language, and the contextual expertise of each reviewer.

To expand the number of Jim Crow laws in the training set, laws containing certain terms were selected for classification by experts. Laws were targeted using keywords related to education and locations, which subject experts identified as topics often associated with Jim Crow laws. Keywords used include: school, district, books, hospital, hotel, restaurant, rail car, and railcar. This targeted selection added another 100 laws to the training set (2 Jim Crow and 98 not Jim Crow).

At this point, the training set comprised 1517 laws (141 Jim Crow, and 1376 not Jim Crow). This set was used to train the model and return probable Jim Crow laws.

The model was trained using 80% of the training set; the remaining 20% was used as a test dataset to assess model precision, i.e. the proportion of laws classified correctly as Jim Crow. Models with each representation were tuned using cross-validation, and performance was assessed using the test dataset. An XGBoost model was selected for prediction (Chen, 2016). To better reflect a given law's probability of being Jim Crow, model output probabilities were calibrated using scikit-learn's CalibratedClassifierCV. Laws with a calibrated probability of 90% or higher were included in the Jim Crow law corpus and laws verified by experts as non-Jim Crow were excluded. The restrictive classification cutoff of 90% was chosen to minimize false positives. A less conservative break point would return more laws labeled as Jim Crow, but would be more likely to include non-Jim Crow laws.

To build the Jim Crow corpus, results from two models were combined: a preliminary model and the final model, which was selected for best performance. A subset of the laws identified by models as likely to be Jim Crow were verified by expert reviewers.

The final model returned 859 laws with at least 90% probability of being Jim Crow. Of these, 72 had been identified as Jim Crow by the earlier model as well but were deemed to be false positives by an expert. These false positives were excluded from the Jim Crow corpus. 118 laws identified by the earlier model and verified by reviewers to be Jim Crow laws were added to the corpus. The final model achieved 84% precision on this subset.

The Jim Crow corpus contains 905 laws. Given the conservative cutoff selected for identification, it is unlikely that the corpus is a comprehensive compilation of all Jim Crow laws enacted during the period of study. 494 laws included in the Jim Crow corpus were verified by one or more experts. 411 of the laws included in the Jim Crow corpus have yet to be reviewed by an expert and likely include some false positives.

For each law in the corpus, the XML metadata indicates whether the law is a Jim Crow law or not, as well as the source used to make this determination. The "jim_crow_source" element for each law indicates "expert" if only identified by an expert, "model" if the law was classified by the model only, or "model and expert" if the law was both identified by the model and verified by an expert.

XML Generation

The corpora were enriched with metadata as XML. Metadata files were merged using a unique identifier, then added to the corpus as XML elements and attributes. Python ElementTree was used to generate the XML. A schema was created that provides the following information about the corpora:

- volume_title: title of the volume, from the Internet Archive metadata
- year_range: year range, from the Internet Archive metadata
- session_name: general assembly session in which the law was ratified
- laws
    - lawtype: type of law (public, private, public-local, or session) manually gathered
    - title: title of the law type section, manually transcribed
- chapter
    - number: chapter number, generated from OCR and data cleaning
- section
    - JPG_link: derived link to the first jpeg image of the section, on the Internet Archive website
    - PDF_link: link to the pdf and specific page on which the law is found on the Internet Archive website, derived from the Internet Archive metadata (leafNum)
    - book_start_page: section start page printed in the original volume (manually transcribed)
    - jim_crow: whether the law is a Jim Crow law: 0=not a Jim Crow law, 1=Jim Crow law.
    - jim_crow_source: specifies how Jim Crow laws were identified (expert, model, or expert and model)
    - jpeg_start_image: the jpeg image of the section start page
    - number: section number, generated from OCR and data cleaning

Errors and Limitations

This ambitious project created a corpus of nearly 300,000 session laws and a corpus of just over 900 Jim Crow laws. There are errors in the corpora resulting from each stage of the process: corpus creation, processing, and analysis.

As reported in the Optical Character Recognition section of this paper, OCR quality was assessed at both the page level and at the word level. Based on our assessment, the words in the corpus were recognized correctly 83.76% of the time, and we expect that at least 94% of the pages do not have significant OCR errors. When splitting the text into chapters and sections, the team found that OCR performed poorly on pages with text in tabular or other unusual formats. Numbers were also frequently recognized incorrectly, especially 3's and 8's. Some sections of text were skipped by OCR, either because of errors associated with header and marginalia removal, or because text was not readable by the software. Missing areas of text that were identified during the chapter splitting process were transcribed by hand, subject to human error.

The process of splitting the corpus into chapters and sections was by far the most time-consuming part of the project. Even though four rounds of cleanup were done, approximately 10% of identified splitting errors remain in the corpus. Volumes with chapters given in Roman numerals or those with chapters listed in the margins had to be split manually, subject to human error. Errors in the original volumes also occur (mis-numbered sections, etc.). Although the vast majority of chapters are made up of individual sections, the volume > chapter > section breakdown used to split the corpus does not fit the variety of ways volumes are organized. One example is that occasionally, chapters were sub-divided with Roman numerals, each with numbered sections. The chapter > section split used for the corpus does not capture any variation from the typical volume organizational structure.

There are also limitations on the analysis of the laws. Identifying Jim Crow laws is not straightforward, even for subject experts. Some laws were used to accomplish racial discrimination without using specific language about race. Some Jim Crow laws were based not on explicit racial language, but on the way the laws were carried out, or how they disproportionately affected people of color. Many of these types of laws cannot be identified with an algorithm. There is also variation in the interpretation of laws. To train the model, subject experts classified laws as either Jim Crow or not Jim Crow. Even on our team, different scholars classified the same laws differently.

Another limitation to our analysis is that most of our expert-labeled Jim Crow Laws were identified by targeted searching, rather than discovered through random sampling. Therefore, our sample is likely biased towards laws easily discovered via keyword or index searches and may not be representative of the variety of Jim Crow laws that exist in the corpus.

The machine learning classification work will continue into the second phase of the project. The team made the decision to spend more time on chapter and section cleanup than analysis because a corpus with as few errors as possible is critical to future analysis on any topic. Additional time will be devoted to assessing a wider variety of models in the second phase of the project. The 90% likelihood cutoff point chosen to identify Jim Crow laws was conservative. It is unlikely that the corpus is a comprehensive listing of North Carolina's Jim Crow laws. The conservative cut off point provided fewer false positives, but fewer Jim Crow laws as well. Finally, 411 of the laws included in the Jim Crow corpus were not reviewed by an expert and likely include some false positives.

Even with these limitations, this project has created the most extensive listing of North Carolina Jim Crow laws ever compiled, and the training set produced from this work will be a valuable contribution for future analysis.

Next Steps

Work for *On the Books* will continue through May 2021, with funding from the ARL Venture Fund Research and development one-time grant. Next steps include further refinement of the products, such as additional clean-up of split laws, improving machine learning models to identify more Jim Crow laws, and expanding analysis from simply identifying the laws to learning more about them. Analyses may include topic modelling of the Jim Crow laws and investigation of temporal and/or geographic patterns associated with the laws. Additionally, future work will facilitate the use of the corpora in the classroom and investigate the potential for using the methods created for *On the Books* to identify Jim Crow legislation in other southern states.

Appendix I.

Volumes included in the corpus:
Internet Archive ID: privatelawsofsta186667nor
Title: Private laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1866/67

Internet Archive ID: publiclawsofstat186667nor
Title: Public laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1866/67

Internet Archive ID: publiclawsofstat1868nort
Title: Public laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1868

Internet Archive ID: privatelawsofsta186869nor
Title: Private laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1868/69

Internet Archive ID: publiclawsofstat186869nor
Title: Public laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1868/69

Internet Archive ID: privatelawsofsta186970nor
Title: Private laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1869/70

Internet Archive ID: publiclawsofstat186970nor
Title: Public laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1869/70

Internet Archive ID: privatelawsofsta187071nor
Title: Private laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1870/71

Internet Archive ID: publiclawsofstat187071nor
Title: Public laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1870/71

Internet Archive ID: privatelawsofsta187172nor
Title: Private laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1871/72

Internet Archive ID: publiclawsofstat187172nor
Title: Public laws of the State of North-Carolina, passed by the General Assembly [serial]
Session: 1871/72

Internet Archive ID: publiclawsresolu187273nor
Title: Public laws and resolutions, together with the private laws, of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1872/73

Internet Archive ID: lawsresolutionso187374nor
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1873/74

Internet Archive ID: lawsresolutionso187475nor
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1874/75

Internet Archive ID: lawsresolutionso187677nor
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1876/77

Internet Archive ID: lawsresolutionso1879nort
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1879

Internet Archive ID: lawsresolutionso1880nort
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1880

Internet Archive ID: lawsresolutionso1881nort
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1881

Internet Archive ID: lawsresolutionso1883nort
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1883

Internet Archive ID: lawsresolutionso1885nort
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1885

Internet Archive ID: lawsresolutionso1887nort
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1887

Internet Archive ID: lawsresolutionso1889nort
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1889

Internet Archive ID: lawsresolutionso1891nort
Title: Laws and resolutions of the State of North Carolina, passed by the General Assembly at its session ... [serial]
Session: 1891

Internet Archive ID: privatelawsofsta1893nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1893

Internet Archive ID: publiclawsresolu1893nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1893

Internet Archive ID: privatelawsofsta1895nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1895

Internet Archive ID: publiclawsresolu1895nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1895

Internet Archive ID: privatelawsofsta1897nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1897

Internet Archive ID: publiclawsresolu1897nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1897

Internet Archive ID: privatelawsofsta1899nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1899

Internet Archive ID: publiclawsresolu1899nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1899

Internet Archive ID: publiclawsresolu1900nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1900

Internet Archive ID: privatelawsofsta1901nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1901

Internet Archive ID: publiclawsresolu1901nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1901

Internet Archive ID: privatelawsofsta1903nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1903

Internet Archive ID: publiclawsresolu1903nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1903

Internet Archive ID: privatelawsofsta1905nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1905

Internet Archive ID: publiclawsresolu1905nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1905

Internet Archive ID: privatelawsofsta1907nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1907

Internet Archive ID: publiclawsresolu1907nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1907

Internet Archive ID: privatelawsofsta1908nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1908 extra

Internet Archive ID: publiclawsresolu1908nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1908 extra

Internet Archive ID: privatelawsofsta1909nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1909

Internet Archive ID: publiclawsresolu1909nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1909

Internet Archive ID: privatelawsofsta1911nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1911

Internet Archive ID: publiclawsresolu1911nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1911

Internet Archive ID: publiclocallawso1911nort
Title: Public local laws of the state of North Carolina passed by the General Assembly [serial]
Session: 1911

Internet Archive ID: privatelawsofsta1913nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1913

Internet Archive ID: publiclawsresolu1913nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1913

Internet Archive ID: publiclocallawso1913nort
Title: Public local laws of the state of North Carolina passed by the General Assembly [serial]
Session: 1913

Internet Archive ID: publiclocallaws1913nort
Title: Public local laws of the State of North Carolina passed by the General Assembly at its session of
Session: 1913 extra

Internet Archive ID: privatelawsofsta1915nort
Title: Private laws of the state of North Carolina passed by the General Assembly at its session of ... [serial]
Session: 1915

Internet Archive ID: publiclawsresolu1915nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1915

Internet Archive ID: publiclocallawso1915nort
Title: Public local laws of the state of North Carolina passed by the General Assembly [serial]
Session: 1915

Internet Archive ID: publiclawsresolu1917nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1917

Internet Archive ID: publiclocallawsp1917nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1917

Internet Archive ID: publiclawsresolu1919nort
Title: Public laws and resolutions of the State of North Carolina [serial] : passed by the General Assembly at its session of ..
Session: 1919

Internet Archive ID: publiclocallawsp1919nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1919

Internet Archive ID: publiclawsresolu1920nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1920 extra

Internet Archive ID: publiclawsresolu1921nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1921

Internet Archive ID: publiclawsresolx1921nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1921 extra

Internet Archive ID: publiclocallawsp1921nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1921 extra

Internet Archive ID: publiclawsresolu1923nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1923

Internet Archive ID: publiclocallawsp1923nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1923

Internet Archive ID: publiclawsresolu1924nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1924 extra

Internet Archive ID: publiclocallawsp1924nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1924 extra

Internet Archive ID: publiclawsresolu1925nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1925

Internet Archive ID: publiclocallawsp1925nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1925

Internet Archive ID: publiclawsresolu1927nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1927

Internet Archive ID: publiclocallawsp1927nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1927

Internet Archive ID: publiclawsresolu1929nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1929

Internet Archive ID: publiclocallawsp1929nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1929

Internet Archive ID: publiclawsresolu1931nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1931

Internet Archive ID: publiclocallawsp1931nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1931

Internet Archive ID: publiclawsresolu1933nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1933

Internet Archive ID: publiclocallawsp1933nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1933

Internet Archive ID: publiclawsresolu1935nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1935

Internet Archive ID: publiclocallawsp1935nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Session: 1935

Internet Archive ID: publiclawsresolu1936nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Sessions: 1936 extra, 1937

Internet Archive ID: publiclawsresolu193839nor
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1938 extra, 1939

Internet Archive ID: publiclocallawsp3839nort
Title: Public-local laws passed by the General Assembly ; Private laws passed by the General Assembly [serial]
Sessions: 1938 extra, 1939

Internet Archive ID: publiclawsresolu1941nort
Title: Public laws and resolutions passed by the General Assembly at its session of ... [serial]
Session: 1941

Internet Archive ID: publiclocallawsp1941nort
Title: Public-local laws and private laws enacted by the General Assembly [serial]
Session: 1941

Internet Archive ID: sessionlawsresol1943nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1943

Internet Archive ID: sessionlawsresol1945nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1945

Internet Archive ID: sessionlawsresol1947nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1947

Internet Archive ID: sessionlawsresol1949nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1949

Internet Archive ID: sessionlawsresol1951nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1951

Internet Archive ID: sessionlawsresol1953nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1953

Internet Archive ID: sessionlawsresol1955nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1955

Internet Archive ID: sessionlaws195657nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1956 extra, 1957

Internet Archive ID: sessionlawsresol1959nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1959

Internet Archive ID: sessionlawsresol1961nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1961

Internet Archive ID: sessionlawsresol1963nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1963

Internet Archive ID: sessionlaws196365nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1963 extra, 1965

Internet Archive ID: sessionlaws19656667nort
Title: Session laws and resolutions passed by the General Assembly [serial]
Session: 1965 extra, 1966 extra, 1967

Bibliography

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016. https://arxiv.org/abs/1603.02754.

Dalwadi, Rucha. 2020. "Analyzing Session Laws of the State of North Carolina: An Automated Approach Using Machine Learning and Natural Language Processing." Master's Paper, University of North Carolina at Chapel Hill. https://doi.org/10.17615/tksc-t217.

Jones, Karen Spärck. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation* 28: 11–21. https://doi.org/10.1.1.115.8343.

Murray, Pauli. 1951. States' Laws on Race and Color: And Appendices Containing International Documents, Federal Laws and Regulations, Local Ordinances and Charts. Cincinnati: Woman's Division of Christian Service, Board of Missions and Church Extension, Methodist Church.

Le, Quoc, and Tomáš Mikolov. 2014. "Distributed Representations of Sentences and Documents." *ICML'14: Proceedings of the 31st International Conference on International Conference on Machine Learning* 32 (2): 1188–1196. http://proceedings.mlr.press/v32/mittelman14.pdf.

Paschal, Richard. forthcoming. *Jim Crow in North Carolina The Legislative Program from 1865 to 1920*. Durham: Carolina Academic Press.