Vegetation History and Archaeobotany
https://doi.org/10.1007/s00334-019-00753-2

ORIGINAL ARTICLE

Discovering accounts of Native American burning within digitized historical documents using information retrieval methods

Stephen J. Tulowiecki[1] · **Scott V. Williams**[1] · **Mary E. Oldendorf**[1]

Received: 9 May 2019 / Accepted: 27 September 2019

Abstract
Historical accounts (ca 17th–19th centuries CE) are valuable for understanding how and where past Native American cultures used fire as a land management tool. Previous research has compiled and interpreted accounts, but methods of compiling them remains less systematic, leaving open the possibility that undiscovered accounts exist. This study uses information retrieval methods to locate accounts of Native American burning within digitized historical documents. Utilizing known accounts from digitized documents, this research develops a model to rank text portions within unread documents based on their predicted relevance. The model used frequencies of key terms and related textual features as predictors, and the presence and absence of accounts within text portions as the dependent variable. Within 121 documents related to western New York State (NYS), USA, this study discovered 40 accounts including 28 in western NYS. Of accounts in western NYS, 12 accounts (describing 21 locations) made explicit connections to Native American burning, were resolvable to town-level or finer resolution, and were not derivative of other texts. To locate all known accounts, the model aided in reducing the amount of text to read to only 0.61% of total. Locations of burning were 0.0 to 16.2 km (median = 5.6 km) from the nearest Native American village area, and 0.1 to 12.5 km (median = 1.5 km) from the nearest trail. This study demonstrates how information retrieval can discover accounts of Native American burning, and suggests that undiscovered historical accounts exist that may advance historical, cultural, and ecological understandings of burning practices.

Keywords Native American · Fire · Information retrieval · Biogeography · North America

Introduction

Past Native American societies of North America used burning as a land management tool for many purposes, such as to create environments favored by game they hunted (e.g. grasslands, forest edges), to ease travel by clearing or thinning forests, and to promote beneficial fire-adapted plant species (Williams 2000; Abrams and Nowacki 2008; Smith 2011). Studying Native American burning practices is important for understanding anthropogenic burning as a component in earth systems (Bowman et al. 2011), assessing human versus natural drivers of forest compositional change over time (Parshall and Foster 2002; McEwan et al. 2011), and comprehending landscapes prior to European or Euro-American arrival (Denevan 1992; Vale 2002). Knowledge of burning is also used to create land management plans, though debate surrounds both its incorporation into conceptualizations of "natural" fire regimes and into ecosystem management (Keeley et al. 2009).

Research has used various data sources and methods to understand and reconstruct patterns of Native American burning. In addition to datasets such as original land survey records, early maps, archaeological records, tree-ring records, and pollen or charcoal records, studies have consulted written accounts (e.g. Fig. 1) within historical documents (Whitney 1996; Brown 2000). Early researchers on the topic (e.g. Day 1953; Stewart 2002) compiled European or Euro-American first-hand accounts of burning and landscapes modified by burning ca. 17th–19th centuries CE. Environmental histories have similarly made use of these accounts (Pyne 1982; Cronon 1983; Whitney 1996). Whitney (1996) compiled a table of 33 accounts in eastern North America, recording attributes such as the year, location, and

Communicated by K. Brown.

✉ Stephen J. Tulowiecki
tulowiecki@geneseo.edu

[1] SUNY Geneseo, 1 College Circle, Geneseo, NY 14454, USA

Published online: 10 October 2019

The origin of the peculiar appearance of these grounds is probably this. The Indians annually, and sometimes oftener, burned such parts of the North American forests, as they found sufficiently dry. In every such case the fuel consists chiefly of the fallen leaves; which are rarely dry enough for an extensive combustion, except on uplands; and on these only when covered with a dry soil. Of this nature, were always the oak, and yellow-pine grounds; which were therefore usually subjected to an annual conflagration. The beech and maple grounds were commonly too wet to be burned. Hence on these grounds the vegetable mould is from six inches to a foot in depth: having been rarely or never consumed by fire; while on the oak and pine grounds it often does not exceed an inch. That this is the effect of fire only, and not of any diversity in the nature of the trees, is evident from the fact, that in moist soils, where the fire cannot penetrate, the mould is as deep on the oak as on the maple grounds. This mould is combustible, and by an intense fire is wholly consumed.

Fig. 1 An example of an account of Native American burning from east of Buffalo, New York State in 1804 (Dwight 1823)

purpose of the burning observed. Williams (2005) compiled a 130-page annotated bibliography of references on Native American burning, a portion of which contained accounts within documents from the 17th–19th centuries. Elsewhere, researchers have compiled databases of indigenous fire use globally. Scherjon et al. (2015) used the electronic Human Relations Area Files database (Yale University 2019), containing historical and ethnographic documents and other sources, to compile 231 accounts of "off-site" indigenous fire use across the globe, including 87 from North America. Scherjon et al. (2015) also recorded attributes of the accounts such as the described purpose of the burning or the time of year when the burning took place. Such accounts have also appeared in land management plans: for example, in a management plan for managed oak openings in western New York State (NYS), Keister (1998) cited accounts of late 18th century Euro-American settlers who witnessed Native American burning in oak openings to justify prescribed burning.

The importance of historical accounts of Native American burning, coupled with limitations of previous research, motivate the development of improved methods to discover additional historical accounts. One presumed limitation of using historical accounts is the manner in which historical accounts are discovered. While previous research often did not explain methods of discovering historical accounts, it is assumed that their discovery was the product of lengthy reading of early historical works, or the use of accounts from texts already cited in previous research. Consequently, research has often utilized the same set of historical accounts. For example, five aforementioned sources (Day 1953; Pyne 1982; Cronon 1983; Whitney 1996; Stewart 2002) all cited Morton's (1883) observations in the 1620s–1630s of burning near Massachusetts Bay, USA, and four of those five sources cited Dwight's (1823) observations in 1804 of burning impacts in western NYS, USA (Fig. 1). Elsewhere, Scherjohn et al. (2015) performed simple searches on a full-text database using terms such as "burn(ing/t)", "charcoal" and "fire". It appears likely that additional historical accounts exist with potential to reveal new insight into Native American burning, including its spatial distribution, and its regional and cultural variations. New methods of discovering such accounts may thereby improve the otherwise "fragmentary" (Williams 2000) documentation of Native American burning.

Advances in information retrieval, alongside machine learning and computational linguistics, present methods for increasing the speed and volume of discovery of historical accounts. One common approach is the "bag-of-words" (Harris 1954) approach: unstructured text is transformed into a document-term matrix (DTM), a table in which a row corresponds to a text portion (e.g. a document, paragraph, or other unit of text), and columns correspond to counts of terms within those text portions (Table 1). Term frequencies serve as predictors (a.k.a. independent variables) in models that predict the relevance of unread text, or that classify unread text using quantitative methods. Previous research has demonstrated success in automated discovery of historical accounts pertinent to an application. For example, Tulowiecki (2018) used information retrieval to efficiently

Table 1 An example of a document-term matrix (DTM) containing just two text portions from Dwight (1823) and five terms

Text portion	"burn"	"county"	"fire"	"forest"	"indian"
"The origin of the peculiar appearance of these grounds is probably this. The Indians annually, and sometimes oftener, burned such parts of the North American forests…"	2	0	4	1	1
"The County of Genesee comprises the whole Western end of the State of New-York. It is bounded on the North by Lake Ontario, on the South by Pennsylvania…"	0	5	0	0	0

The first text portion is beginning of the portion in Fig. 1. Entire text portions are not shown, and therefore key term counts do not match the number of terms shown in the "Text" column

discover hundreds of forest compositional descriptions at Euro-American arrival within 18th–20th century county histories, developing a ranking model that predicted the relevance of text based largely on the number of unique tree species listed in each text portion (e.g. a paragraph) as a predictor.

The purpose of this study is to test whether information retrieval methods can discover accounts of Native American burning (hereafter "accounts") within digitized historical documents. This study trains a model that ranks portions of text based on their probability of containing accounts, using key term frequencies and other features of the text as predictors. This study addresses the above limitations by developing methods to discover accounts more efficiently and to discover unstudied accounts.

Materials and methods

This study first developed and tested a model to rank portions of text from digitized historical documents by their probability of containing an account. This study then mapped the discovered accounts and recorded their characteristics. Tasks involving geographic information systems (GIS) software were performed using ArcGIS 10.5 (Esri 2017), and tasks involving R statistical computing were performed using R version 3.5.1 (R Core Team 2018) and RStudio version 1.0.153 (RStudio Team 2016).

Developing the ranking model

A ranking model was developed based on the relationship between the presence and absence of accounts (e.g. Fig. 1) within text portions (the dependent variable), and features of text portions such as the presence or frequency of certain key terms (the predictors). The model was trained using text portions from digitized historical accounts known to contain accounts. Once developed, the model's predictive ability was tested by using it to discover accounts in an independent collection of documents. The following four sub-sections describe steps taken to develop and test this model.

Discovering terms related to accounts

First sought was an understanding of key terms most associated with known accounts. A list of known accounts of Native American burning cited in previous literature was created. Two works were examined that quoted several commonly-referenced accounts: an article by Day (1953) and a book by Stewart (2002), two early major works on Native American burning and its impacts (Lewis and Anderson 2002). Accounts from additional documents known to the authors of this study were also included in the list. Accounts either explicitly described Native American burning through first-hand observation, or speculated that landscape conditions (e.g. prairies, oak savannas, or open forests) were attributable to burning. A few accounts were included that described landscape conditions suggestive of burning, but did not explicitly attribute them to fire. The accounts were from early explorers (ca 17th–18th centuries), early travellers (ca 18th–19th centuries), writers of histories such as county histories (ca 19th century), and early researchers on Native American burning (ca 19th–20th centuries). Accounts were collected from a wide geographic area to consider any regional variations in terminology among the accounts.

A total of 129 accounts spread over 269 text portions were collected from 105 historical documents. In this study, "text portion" refers to the text between double-line breaks in the .txt-format documents. A text portion was typically a paragraph; however, due to factors such as page breaks, running titles, and errors that occur when converting scanned documents to .txt-format files, text portions were sometimes shorter than a paragraph. Since this study later developed a model to predict whether a text portion contained part or all of an account by breaking a new .txt-format document into text portions using double-line breaks, it was similarly used for this part of the study.

Whereas some locations of burning could be pinpointed to town-resolution (roughly within a 10×10 km area) or finer, some accounts were spatially imprecise and described burning somewhere within large geographic areas. Figure 2 shows the approximate locations and positional uncertainty of the accounts collected. For later model development steps, historical documents producing the accounts were downloaded. Text (.txt)-format versions of 100 of the 105 were locatable and downloaded from online collections, typically from the Internet Archive (2019).

To determine the most common terms in accounts, a script was developed in R using the "tm" (Feinerer et al. 2008; Feinerer and Hornik 2018) and "dplyr" (Wickham et al. 2018) packages. The script loaded the list of accounts and performed tasks common to computational analysis of text: it removed all non-letter characters, converted all letters to lowercase, removed common stopwords (e.g. "at", "is", "on"…) and stemmed the remaining words (e.g. "burned", "burning", etc. were converted to "burn"). A DTM was then created to determine which words appeared most frequently throughout all text portions. The script was developed to produce a DTM with words appearing in $> 1\%$ of text portions, totalling 1,100 + terms. The terms were then examined to select those that were believed to be the most predictive of accounts. This process involved determining which of the terms were more frequent in accounts than in other text portions within the historical documents overall. To aid this process, another R script determined the frequencies of the

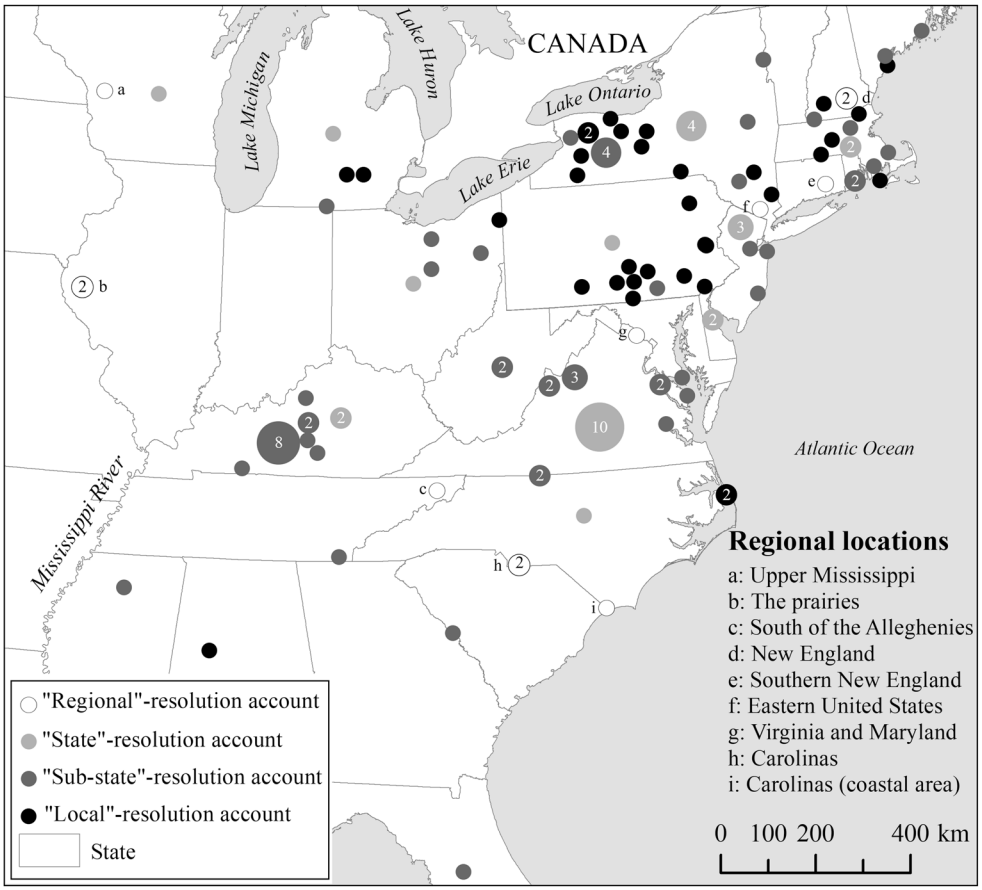

Fig. 2 A map depicting locations of accounts used during model development. Symbol size and numeric labels indicate the number of accounts describing a given location. Locations of accounts are ambiguous and the resolution describes their positional uncertainty. "Regional" = account describes a general region, or a location somewhere within an area larger than multiple US states (e.g. "Eastern US"). "State" = account describes a location somewhere within a state (e.g. "New York"). "Sub-state" = account describes a location somewhere within a region smaller than a state (e.g. "western New York"). "Local" = account describes a location somewhere within a localized area (e.g. "Town of Groveland")

1,100 + terms in all text portions within the downloaded historical documents, to compare with the term frequencies in just text portions comprising the collected accounts. For example, the word "undergrowth" appeared in 4.46% of text portions that comprised accounts but only 0.02% of all text portions from the downloaded training texts, more than 200 times more frequent in text portions from accounts. After this process, 159 terms remained (Table 2).

Also created was a list of words that were purportedly negatively associated with accounts. Specifically sought were terms that appeared in text portions that did not contain accounts, but that contained at least one fire-related term and one Native American-related term (Table 2). These terms were sought for their presumed importance in differentiating text portions with accounts from other accounts of Native American fire use (e.g. cooking, ceremonial, warfare…) or other topics altogether (e.g. war). Another R script selected text portions from the downloaded texts with at least one fire-related term and one Native American-related term, and then created a DTM. The list of terms in this DTM was then cross-referenced with the previous list of 1,100 + terms from the text portions comprising accounts, to find terms that did not occur in this list. A total of 683 terms not used in any of the collected accounts were compiled in this manner.

Developing predictors

Using the terms discovered above, 58 predictors were created and considered for inclusion in the ranking model. A full list of predictors created and considered is presented in Table 3. Examples of predictors included the total number of appearances of specific terms, the percentage of total terms that were a specific term, and the number of unique terms from a category of terms. Due to correlation between some predictors, five were removed from future model development, leaving 53 predictors. Correlation was determined

Table 2 Terms and categories for the development of predictors in the ranking model

Category	Terms
Fire-related	Burn, charcoal, conflagr, consum, fire, flame, fuel, kindl, light, set, smoke, spread
Fire-related, "other"	Blaze, combust, ignit, incendiary, lit, inferno, scorch, torch
Hunting-related	Anim, bear, beast, buffalo, deer, elk, game, graze, herd, hunt, kill, pastur, shoot, turkey
Landscape-related	Area, bare, bark, barren, beauti, brush, bush, charact, clear, countri, cover, creek, dens, destitut, devoid, district, dri, earth, elev, extent, fertil, field, flat, forest, grass, gravel, green, ground, grove, herb, hill, land, level, meadow, mountain, natur, open, park, place, plain, plant, prairi, region, river, sand, savanna, scarc, scatter, shrub, space, speci, stand, stream, surfac, swamp, tall, thick, thin, timber, timberless, tract, tree, treeless, underbrush, undergrowth, underwood, upland, valley, vast, veget, wet, wild, wood
Native American-related	Aborigin, indian, inhabit, savag, settl, settlement, tribe
Native American-related, "other"	Injun, native, occup, people, race, redman, resid, squaw
Time-related terms	Annual, autumn, fall, season, spring, summer, time
Xerophytic taxa-related	Chestnut, hickori, oak, pine, walnut
Miscellaneous terms	Abound, abund, appear, common, corn, crop, cultiv, custom, dead, destroy, distanc, drive, driven, easili, extend, extens, found, free, frequent, great, grow, habit, height, larg, leagu, leav, length, littl, low, mani, mile, part, plantat, quantiti, small, soil, travel, varieti, whole, wind, young
"A"-list	Aborigin, anim, barren, beauti, brush, buffalo, burn, bush, charcoal, chestnut, clear, conflagr, consum, countri, cover, custom, deer, destitut, destroy, devoid, dri, elk, field, fire, flame, forest, game, grass, ground, grove, grow, herb, hickori, hunt, indian, inhabit, kindl, land, leav, level, meadow, oak, open, pastur, pine, place, plain, prairi, region, savag, scatter, set, shrub, smoke, soil, space, surfac, thick, thin, timber, timberless, travel, tree, treeless, tribe, underbrush, undergrowth, underwood, upland, veget, wild, wood
"B"-list	Abound, abund, annual, appear, area, autumn, bare, bark, bear, beast, charact, common, corn, creek, crop, cultiv, dead, dens, distanc, district, drive, driven, earth, easili, elev, extend, extens, extent, fall, fertil, flat, found, free, frequent, fuel, gravel, graze, great, green, habit, height, herd, hill, kill, larg, leagu, length, light, littl, low, mani, mile, mountain, natur, park, part, plant, plantat, quantiti, river, sand, savanna, scarc, season, settl, settlement, shoot, small, speci, spread, spring, stand, stream, summer, swamp, tall, time, tract, turkey, valley, varieti, vast, walnut, wet, whole, wind, young
Negatively associated terms	Church, enemi, born, busi, meet…

Terms shown are stemmed; e.g. "burned", "burning", etc. were reduced to "burn". Note that the categories are not mutually-exclusive; e.g. the word "burn" appears in two categories. We acknowledge that some historical terms for Native Americans that appear are racially insensitive

using the training data (next section). Further description of how predictors were created and chosen is provided below.

To create some predictors, the 159 terms assumed to be positively correlated (Table 2) were classified into seven categories based on their meaning in accounts. Due to the anticipated importance of terms related to Native Americans and fire in predicting the presence of accounts within text portions, 16 additional terms related to these topics were also added by consulting thesaurus entries for "Indian", "fire", and related terms. To create different subsets of terms to derive other predictors, an "A"-list and a "B"-list were created from the 159 terms (Table 2). The "A"-list included terms believed to be the most positively associated with accounts because they appeared in many accounts, and/or because their frequency in accounts was much higher than their background frequency in the digitized historical documents. For example, the term "fire" was often used in the collected accounts and was also more frequent in relevant accounts compared to non-relevant text portions, so it was included in the "A"-list. As another example, the word "great" was often used in the collected accounts but was only slightly more frequent in relevant accounts compared to non-relevant text portions, so it was included in the "B"-list. Term categories described above and in Tables 2 and 3 are not mutually exclusive; the intention of creating different subsets of terms was to create and experiment with various predictors to consider in the model, in order to create the most predictive ranking model.

Training the model

A model was trained that related the presence and absence of accounts to the 53 predictors. The training data were comprised of text portions containing presences (i.e. accounts) and absences (i.e. no accounts). The presences were 111 text portions that explicitly described Native American burning or speculated that burning caused observed landscape conditions, from the 129 accounts collected and described previously. Selected for absences were over 11,000 random text portions from the downloaded historical documents. The number of absences selected from a document was proportional to its length.

Boosted regression trees (BRT; Friedman 2001, 2002; Elith et al. 2008) were used to develop a ranking model.

Table 3 Predictors created for ranking models. Predictor values were calculated per text portion. Refer to Table 2 for further understanding of predictors explained here

Predictor
Percentage of total terms (minus stopwords) that are…
Fire-related
Fire-related, "other"
Hunting-related
Landscape-related
Native American-related
Native American-related, "other"
Time-related
Xerophytic taxa-related
Miscellaneous
"A"-list
"B"-list
Presumed to be negatively associated
Number of unique terms that are…
Fire-related
Fire-related, "other"
Hunting-related
Landscape-related
Native American-related
Native American-related, "other"
Time-related
Xerophytic taxa-related
Miscellaneous
"A"-list
"B"-list
Presumed to be negatively associated
Total number of appearances of the term…
"burn"
"charcoal"
"conflagr"
"consum"
"fire"
"flame"
"fuel"
"kindl"
"light"
"set"
"smoke"
"spread"
"blaze"
"combust"
"ignit"
"incendiari"
"lit"
"scorch"
"torch"
"aborigin"
"indian"
"inhabit"
"savag"
"settl"
"settlement"
"tribe"
"nativ"
"occup"
"peopl"
"race"
"redman"
"resid"
"squaw"
Total number of terms (minus stopwords)

A machine-learning technique, BRT works by fitting a regression tree upon a random subset of samples. Each split in the regression tree is made by choosing the value from a predictor that best minimizes predictive deviance in the subset, until the maximum number of splits (the "tree complexity", or *tc*) is reached. *tc* equals the number of allowed predictor interactions (and can be parameterized by training several BRT models and examining which *tc* value produces the best predictions during cross-validation). A subsequent tree is then fitted using the residuals that occur when the prior regression tree(s) is/are applied to predict upon a new random subset. The "learning rate", or *lr*, determines how much each tree contributes to the final BRT model. The final BRT model can be conceptualized as a model in which each regression tree is a model term (Elith et al. 2008). Tulowiecki (2018) used BRT to predict the presence of descriptions of original forest composition within text portions from county histories.

BRT models were trained in R using the "dismo" package created by Hijmans et al. (2013), and using cross-validation methods summarized in Elith et al. (2008), with *lr* values that resulted in between 1,000 and 5,000 trees (number of trees, or *nt*). Different *lr* values from 0.001 to 0.005 and *tc* values from 1 to 10 were tried, and the final parameters were selected to achieve the lowest predictive deviance during cross-validation. Additional cross-validation methods available in the "dismo" R package (Hijmans et al. 2013) were used to remove predictors from the initial model that did not appreciably improve the model. The probability value outputted from the final model was interpreted as a relevance score for each text portion. The importance of predictors was assessed using the relative variable importance measure. Partial dependence plots

were used to characterize the relationship between predictors and the relevance of a text portion.

Testing the model

We applied the BRT model to discover accounts within a set of historical documents associated with western NYS (approximately 28,000 km^2), USA, which currently embraces part or all of 16 counties. This region was chosen because previous research compiled accounts from documents associated with the region (Tulowiecki et al. 2019), and these accounts provide a unique opportunity to test the ability of this study's model to discover known accounts. Moreover, the region was home to numerous Native American cultures up to and beyond the late 18th century, most notably the Seneca of the Haudenosaunee (Iroquois) Confederacy (Snow 1996; Engelbrecht 2003). To test the model, 121 historical documents were used that Tulowiecki et al. (2019) examined for accounts of oak savannas and Native American burning in western NYS. They discovered 25 accounts of early burning (not all attributed directly to Native Americans) using a Python script that inputted .txt-format documents and outputted text portions containing key terms they specified. They applied various rules to sort and examine the outputted text portions based on key term presences and frequencies, but did not develop a ranking model. The 121 documents were comprised of 80 histories (e.g. county or town histories), 22 journals (e.g. by early travellers or military officers), 10 anthropological or ethnographical texts, and 9 gazetteers or related texts. A few documents contained accounts also used for developing the ranking model, but > 95% of the documents were not used for training the model in this study.

R scripts were written to (1) split each document into text portions (defined as the text between double-line breaks in the .txt-format documents), (2) calculate values for the predictors for each text portion, and (3) apply the model to generate the relevance score for each text portion based on predictor values. The scripts inputted the .txt-format documents and outputted a spreadsheet containing the relevance score, the text portion, and the document file name. Sorted in order from highest to lowest predicted relevance, text portions were read to evaluate whether they contained an account. The top 500 text portions by predicted relevance were read to discover accounts; these included accounts discovered by Tulowiecki et al. (2019), accounts not discovered previously, and accounts both inside and outside western NYS. Accounts discovered by Tulowiecki et al. (2019) were located in the ranked output to understand how many text portions would need to be read to discover those accounts. In cases where text portions were suggestive of Native American burning but required further reading of the source document to make a determination, those documents were read to ascertain whether the account was relevant. For model validation, a text portion was marked as "relevant" if it provided textual clues that led to the discovery of an account, even if the text portion did not explicitly provide the necessary information to make that determination.

To understand how many top-ranked text portions would need to be read within each document to discover each account, the within-document ranks of text portions leading to a relevant account were also recorded. The outputted spreadsheet was sorted first by document name and then by predicted relevance of each text portion (from highest to lowest). Examining text portions for each document, the rank order of text portions that comprised relevant accounts was recorded; e.g. if a text portion from a relevant account was the fifth highest-ranked text portion from the document, its rank was 5. Only documents with at least one account were included in this analysis.

Various measures assessed the ranking performance of the model, such as the total number of relevant accounts discovered, and the percentage of accounts discovered out of all text portions read. Charts were also created to visualize the ability of the model to rank the text portions. x-axes were the cumulative number of words or text portions read, whereas y-axes included cumulative number of relevant text portions or unique accounts discovered. Similar charts were created by applying the model to the training data to assess model fit. Collectively these charts focused on the model's ability to lead to discovery of unique accounts rather than to discover individual relevant text portions (since multiple text portions often comprised one account), because ultimately the goal was to discover unique accounts.

Mapping and recording attributes of historical accounts

Using GIS software, accounts were mapped as point locations based on locational information presented within them. Typically, the centroid of the described area served as the point representation of the account; for example, if an account described a burned landscape within a town, the centroid of the town was used. Attributes recorded were: the finest resolution to which the account could be located (e.g. town); whether the account appeared to be derivative of an earlier document; whether Native American burning was explicitly observed or linked to the observed landscape conditions; and if applicable the purpose of the burning.

To understand additional geographic characteristics of the accounts, the distance from the accounts to ca 18th century Native American trails and village locations was estimated using GIS software. Morgan's map (1901) of major trails was georeferenced and traced, a source used previously to understand Haudenosaunee settlement patterns (Jones 2010). These trail locations were viewed only as approximations

due to positional error and generalization in the map. While additional trails likely existed within the study area, later analyses suggested that digitizing other trails from additional sources would not appreciably change minimum distances from account points to trails. To map village locations, GIS data on the locations of village areas were acquired from Tulowiecki et al. (2019), which they compiled from six sources: Cappon (1976), Grumet (1995), Hays and Post (1999), Jennings and Fenton (1995), Morgan (1901), and Parker (1920). Rather than represent villages as point locations, Tulowiecki et al. (2019) used generalized polygons to represent areas with one or more villages.

Results

Ranking model performance

This study's ranking model demonstrated a strong ability to rank text portions by relevance, and to reduce the amount of text to examine in order to discover relevant accounts. Within the 500 most highly-ranked portions, 59 (11.8%) relevant text portions were discovered, leading to the discovery of 40 unique accounts (some accounts were made up of multiple text portions). Included in the top 500 were 22 out of 25 (88.0%) unique accounts discovered by Tulowiecki et al. (2019) in western NYS, plus 6 accounts in western NYS not discovered in that study. Over 50% of the accounts compiled by Tulowiecki et al. (2019) were discovered in the top 39 text portions. To locate all accounts discovered in this study and in Tulowiecki et al. (2019), 3,664 text portions would need to be read, equalling 0.61% of text from the 121 documents searched.

Relevant accounts were concentrated towards the top of the ordered list of text portions from all documents combined. Figure 3 summarizes the model's performance by plotting cumulative text read (when reading from highest to lowest predicted relevance) versus cumulative number of accounts discovered, when examining text portions from all documents combined. Figure 4 summarizes how many relevant text portions were discovered as a percentage of cumulative text portions read. For example, the top 25 portions contained 12 relevant text portions (48.0%) leading to the discovery of 10 unique accounts (Fig. 4). As more text portions were examined, fewer relevant accounts were discovered. For instance, more relevant text portions were located in the 100 most highly-ranked text portions than in text portions ranked 101 through 500. Despite overall model success, comparing

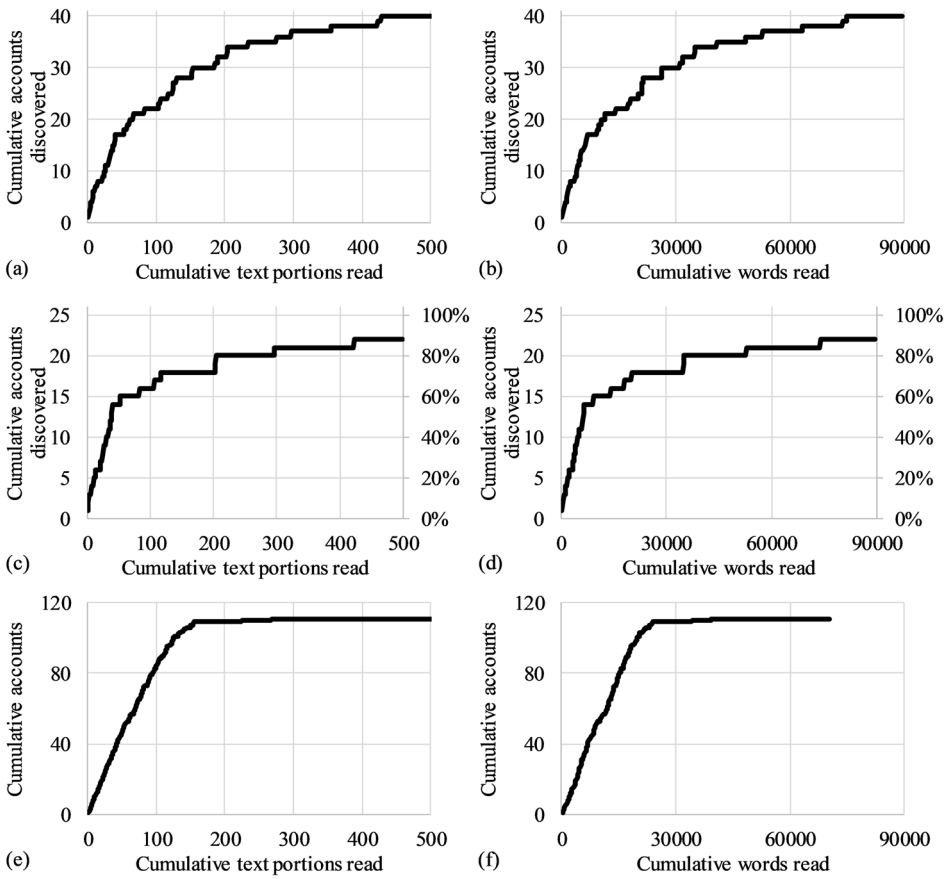

Fig. 3 Total accounts discovered versus **a** cumulative text portions read and **b** cumulative words read; and accounts found by Tulowiecki et al. (in press) versus **c** cumulative text portions read and **d** cumulative words read. For **e** and **f**, the model was applied to the training data and the text portions were similarly ranked. Total words in these charts include stopwords, in order to represent the actual total number of words read

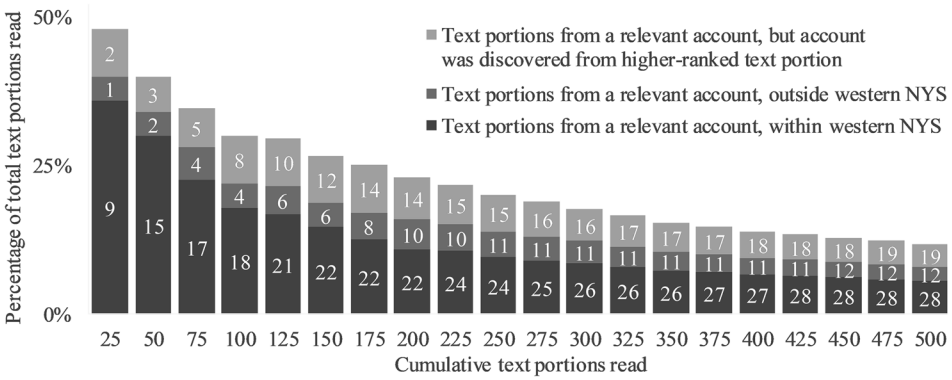

Fig. 4 Cumulative text portions read, versus the percentage of text portions that contained relevant accounts. Labels indicate raw text portion totals. *NYS* New York State

model performance on training versus test data suggests some model overfitting to the training data (Fig. 3): the model better ranked text portions when applied to the training data (Fig. 3e–f) in comparison to the test data (Fig. 3a–d).

Text portions from relevant accounts were also ranked highly within individual documents. Within-document ranks of text portions comprising a relevant account ranged from 1 to 121, with a median of 2. Note that these ranks were for the highest-ranked text portion from each relevant account, as some accounts were comprised of multiple text portions. Out of 43 accounts (40 in the top 500, plus 3 in Tulowiecki et al. (2019) not in the top 500), 38 (86.4%) were represented in the top 10 of text portions within their respective documents, of which 19 (43.2%) were the top-ranked text portion.

Variable selection procedures during the development of the ranking model chose just 13 predictors out of the 58 considered. Table 4 summarizes importance measures for predictors in the final ranking model; the top five most important predictors accounted for 64.5% of variable importance. The three most important predictors were all fire-related: (1) total number of appearances of the term "burn"; (2) total number of appearances of the term "fire"; and (3) percentage of total terms (minus stopwords) that are fire-related. Predictor relationships with the relevance scores were generally as expected, as judged by partial dependence plots; Table 4 also provides a summary of predictor values associated with higher relevance scores.

Characteristics of accounts

This paragraph describes characteristics of the 40 accounts discovered in the top 500 highest-ranked text portions. Overall, accounts varied in the quality of their content and their positional uncertainty. Since this study focused on historical documents related to western NYS, most (28 out of 40) accounts pertained to locations in this region (Fig. 5). A few accounts appeared to have been derivative of earlier accounts, but most (22 of 28 in western NYS, and all 12 outside of

Table 4 Predictors selected by the final boosted regression trees (BRT) model and their relative importance

Predictor	Relative importance (%)	Relevance score is generally higher when…
Total number of appearances of the term "burn"	23.9	≥ 1
Total number of appearances of the term "fire"	12.1	≥ 1
Percentage of total terms (minus stopwords) that are fire-related	12.1	> 0%
Percentage of total terms (minus stopwords) that are presumed to be negatively associated	9.6	< 15%
Percentage of total terms (minus stopwords) that are "A"-list	6.8	> 12%
Percentage of total terms (minus stopwords) that are landscape-related	6.2	> 5%
Number of unique terms that are hunting-related	6.0	≥ 2
Percentage of total terms (minus stopwords) that are Native-American related	5.9	> 0%
Total number of appearances of the term "indian"	5.0	≥ 1
Number of unique terms that are landscape-related	4.1	≥ 4
Percentage of total terms (minus stopwords) that are miscellaneous	3.9	> 2%
Percentage of total terms (minus stopwords) that are "B"-list	2.2	> 2%
Total number of terms (minus stopwords)	2.0	≥ 10

Values do not sum to 100 due to rounding. Also shown are predictor values generally associated with an elevated relevance score for text portions, as judged from partial dependence plots. Refer to Tables 2 and 3 for predictor explanations

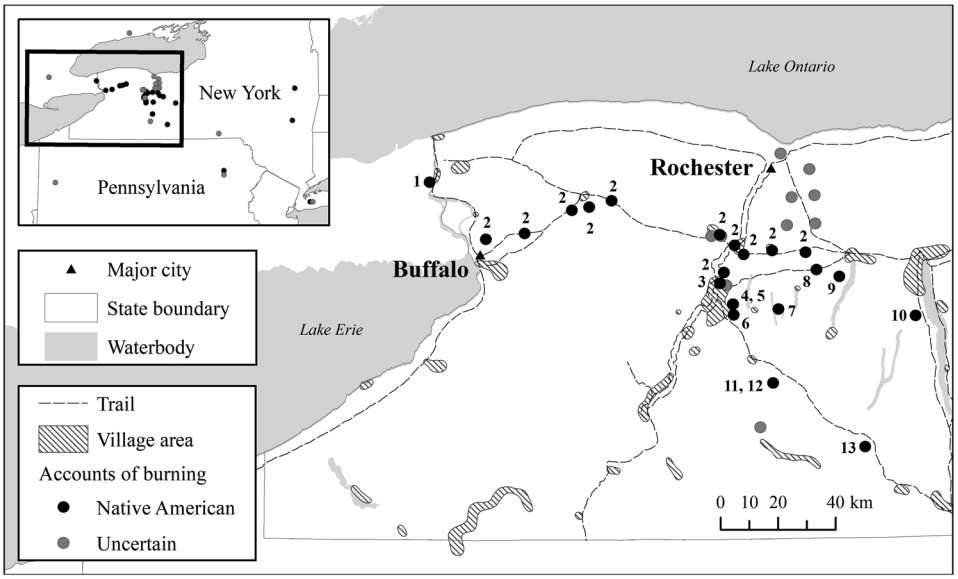

Fig. 5 A map of accounts in western New York State and beyond. Only original, finer-resolution accounts are mapped. Number labels correspond to "higher-quality" accounts (12 within top 500 text portions, plus 1 outside the top 500; see "Results" section) within the study area, excerpts of which are provided in Table 5. All points symbolizing accounts are visible in the inset map except for two

western NYS) appeared to provide original information. Most of the original accounts (15 of 22 in western NYS, and 4 of 12 outside of western NYS) explicitly described Native American burning or speculated that it was the cause of observed landscape conditions; remaining accounts were ambiguous and possibly described Native American or early Euro-American land burning. A high majority of accounts (19 of 22 in western NYS, and 11 of 12 outside of western NYS) were of finer spatial resolution, generally resolvable to town resolution (within a 10 × 10 km area) or finer. Summarizing the accounts differently, 12 of 28 in western NYS accounts, and 4 of 12 outside of western NYS, were considered "higher-quality"—they were original accounts, made explicit connections to Native American burning, were resolvable to finer resolution, and were not derivative of other documents. Excerpts from these higher-quality accounts are provided in Table 5.

Roughly one account was found out of every three documents comprising the test data. The number of accounts within a document ranged from 0 to 4 accounts. Histories produced the highest total of accounts, with 80 histories producing 27 accounts (of which 11 were higher-quality). By percentage, travel narratives produced the most accounts, with 22 narratives producing 12 accounts (of which 4 were higher-quality). 9 gazetteers or other documents produced 3 accounts (of which 2 were higher-quality), and 10 archaeological or ethnographical documents produced 1 account.

Within western NYS, accounts showed spatial clustering and relationships with Native American settlement patterns (Fig. 5). Distance from nearest Native American village areas to points representing locations of burning ranged from 0.0 to 16.2 km with a median of 5.6 km. Distances from nearest Native American trails to points representing locations of burning ranged from 0.1 to 12.5 km with a median of 1.5 km. These distance calculations included only unique locations described in higher-quality accounts within western NYS from the top 500 text portions, plus one location described by a higher-quality account outside of the top 500. Considering only original (non-derivative) accounts, the accounts described 31 unique locations within western NYS and 12 unique locations outside of western NYS, although some accounts likely used different locational descriptions for the same observed geographic area. Of the higher-quality accounts, 12 in western NYS described 21 locations of burning, and the 4 outside western NYS described 4 locations. All accounts discovered in this study described a single location, except for one account that listed 11 locations and another that listed 6 locations. Some locations were described by multiple accounts.

Of the 12 higher-quality western NYS accounts, 10 ascribed one or more purposes to the Native American burning. Most higher-quality accounts linked Native American burning to hunting practices—8 of 12 suggested that the purpose of burning was to encourage browse and/or clear land to create habitat for deer. Few other purposes were ascribed to the burning: 2 linked Native American burning to general land clearance; 1 linked burning to clearing for horticultural purposes; and 1 account linked burning to the extermination of rattlesnakes. Of the 4 higher-quality accounts outside of western NYS, 2 linked burning to hunting practices and 1 linked burning to general land clearance.

Discussion

The potential of information retrieval to discover historical accounts of burning

This study demonstrates the value of information retrieval methods for locating historical information within digitized

Table 5 Excerpts from higher-quality accounts within the study area

ID	Source	Excerpt from account
1	Turner (1849)	"[Rattle-snakes] were so numerous at one time, at their principal den below [Niagara] Falls, that the Tuscarora Indians could not safely occupy a favorite fishing ground there. They extirpated them in great numbers, by setting fire to the dry leaves, burning over the steep bank…"
2	Dwight (1823)	"These grounds are also termed Openings; as being in a great degree destitute of forests…These grounds are of a singular, and interesting appearance. The trees, growing on them, are almost universally oaks… The origin of the peculiar appearance of these grounds is probably this. The Indians annually, and sometimes oftener, burned such parts of the North American forests, as they found sufficiently dry."
3	Minard (1896)	"The open flats were covered with a luxuriant growth of grass…In places this grass was burned off, exposing a soil, which, subjected to the manipulations of the rude husbandry of Indian women, laughed with a bountiful harvest…"
4	John M'Kay, in Turner (1849)	"Among the early events that now occur to me, was the firing of lands by the Indians for the purpose of taking game. It was in 1795. The Indians to the number of at least five hundred assembled. At 12 o'clock in the day, they set a train of fire which enclosed an area of about seven miles square, of the oak openings between the Canascraga and Conesus Lake."
5	Doty (1905)	"The pioneers found the surface of the town everywhere diversified with clusters of fine trees, free from undergrowth, with intervals of natural openings. The fires periodically kindled by the Indians had destroyed the leaves and bushes and in a great measure the fallen and decaying wood, so that it presented the appearance of a succession of groves…"
6	Samuel Magee, in Doty (1905)	"What is now called Groveland hill was at first considered very poor land. Many portions were scatteringly covered with chestnut and the different kinds of oak, and some places were destitute of timber altogether. The openings grew up to a tall red grass which was burned over every fall by the Indians. In some parts of the timbered lands would be found an undergrowth of whortleberry and other bushes…"
7	Waite (1883)	"The hill that bounds the eastern shore [of Hemlock Lake] is called both Ball and Bald, the former, from being a pretty true segment of a circle some thirteen or more miles in diameter, and the latter from its bald appearance in a very early day, caused by the frequent fires of the Senecas."
8	McIntosh (1878)	"A small Indian village was at one time located on the rise of land northeast from Baptist Hill, on land now owned by George Andrews. The land throughout this country presented unmistakable evidences of having been frequently burned over by the Indians…The aboriginals undoubtedly resorted to this method to retain the game in the vicinity of their homes"
9	McIntosh (1878)	"District No. 5 contains the village of Cheshire…The lands of this district lie in ridges; hills rise above hills, and in the valleys was marshy land, covered with a heavy growth of oak, poplar, and butternut; on the highlands the forest-trees were fewer and smaller, and hence more easily cleared. The Indians had burned the woods annually, and, caring nothing for the trees, the fresh herbage, inducing the presence of deer, was to them of more account."
10	Cleveland (1873)	"It was a country for the most part very heavily wooded, a few ridges forming exceptions, where it is said the Indians had repeatedly burned the land over, for the double purpose of securing open spaces in the forest, and furnishing by the new growth the food most eagerly sought for by the deer and elk…The land for some distance east and northeast of Penn Yan was of this character. That the timber was dwarfish and scattering, was evidently due to some other cause than lack of fertility in the soil."
11	Clayton (1879)	"In the north part of the town, to the east of Stony Brook, was originally a high sandy plain, covered with a light growth of oak and yellow pine, which had been annually burned over by the Indians to make a hunting-ground. When the first settlers came there were about 1000 acres of this so open it could be seen through, and nearly level, some of the surrounding hills being also quite bare."
12	Roberts (1891)	"In regard to the poverty of the soil of those yellow pine plains, it is said to have been caused by the Indians annually burning the leaves for a long series of years. The surrounding hills and valleys were their favorite hunting-grounds, those forests abounded with the red deer and other game."
13	Maude (1826)	"After passing Mud Creek, the road, following on its N. side the course of the Conhocton, was tolerably good; here the Timber was principally Scrub Oak, intermixed with Yellow Scrub Pine: this degeneracy of the wood is owing to its being annually burnt by the Indians; the destructive mode of clearing a passage through the woods, and rousing the game, is now put a stop to, nothing being more destructive to the soil…"

ID numbers correspond to account locations that are mapped in Fig. 5

documents on past fire use. It manifests the feasibility of "big data" approaches towards improving the historical record of anthropogenic burning by increasing the efficiency of discovering historical accounts. In particular, results suggest that future efforts should focus on histories and travel narratives for accounts of Native American burning—histories are more common and produced high amounts of accounts, whereas travel narratives provided more accounts per

document. Coupled with existing bibliographies of historical documents, this study suggests that numerous accounts may exist for future discovery in these sources. A total of 80 histories searched in this study produced 27 accounts (11 higher-quality) of Native American burning, but Filby (1985) estimated that 5,000 county histories alone exist in the United States. A total of 22 travel narratives searched in this study produced 12 accounts (4 higher-quality), but McKinstry (1997) listed 406 narratives across the United States, many from the 18th–19th centuries. Documents that were searched in this study were downloaded from the Internet Archive, which contains approximately 20 million texts (Internet Archive 2019).

This study's ranking model

This study produced an effective ranking model for discovering historical accounts of Native American burning. The importance of predictors related to lists of key terms from known accounts (Tables 2, 3 and 4) demonstrate that the training samples and texts were useful, and the model was mostly transferable to other texts. To our knowledge, few to no studies have used similar methods to discover historical accounts of landscape conditions or land uses, with the exception of Tulowiecki's (2018) BRT model that discovered accounts of forest composition in documents based on the frequency of forest compositional terms (e.g. "maple", "oak"). Though viewed as highly useful, this study's model was likely not as discerning at Tulowiecki's (2018), due to the more common and/or ambiguous nature of key terms (e.g. "burn", "Indian") related to accounts in this study, as well as the rarity of such accounts.

Though difficult to ascertain the true recall (i.e. percentage of all accounts that were found) of the ranking model due to the large volume of text to search, the results showed that reading the top 500 text portions (i.e. from all documents) likely led to the discovery of a high majority of accounts. On average just four text portions per document would need to be examined to locate a high majority of accounts. Moreover, when the ranking model was applied to documents individually, many text portions comprising a relevant account were the top-ranked portion or within the top 10 of ranked portions. Fewer relevant accounts were discovered with a decrease in relevance score, but text portions on other topics related to fire were still ranked highly, including early Euro-American land burning and other Native American uses of fire (e.g. cooking, heating longhouses).

Four limitations and future considerations are highlighted. First, this study utilized a machine-learning technique to develop a ranking model. It is possible that simpler rule-based methods could locate accounts simply by isolating text portions that meet basic criteria (e.g. Table 4, final column), though it may lead to reading more text. Conversely, more complex methods from computational linguistics or information retrieval may improve ranking performance, such as those reviewed in Khan et al. (2010) and Dalal and Zaveri (2011). Second, creating a ranking model should be viewed as a process whereby newly-discovered accounts, along with highly-scored but irrelevant accounts, form new training data and yield insight into new predictors to include in an updated, improved model. Related to this point, we acknowledge that decisions made to select key terms and create predictor variables (Tables 2 and 3) were somewhat subjective and that results may be sensitive to those decisions. Third, future studies should be aware of optical character recognition (OCR) errors, which occur when scanned or photographed documents are converted automatically into .txt-format files (Alex et al. 2012). OCR errors can negatively impact quantitative methods that involve analysis of converted texts (Hill and Hengchen 2019), by incorrectly converting a term into .txt-format and in turn leading to incorrect totals of key terms. Though OCR errors were present in this study's documents, investigating the training samples suggest that these errors had a minimal impact on our model. Fourth, additional modelling decisions could improve the ranking ability of future models; for example, worth exploring is whether varying the size of text portions (e.g. using sentences or entire pages of text) as analytical units to count key terms that form predictors would produce better models.

Patterns and uses of Native American burning revealed by historical accounts

This study furthermore demonstrated how information retrieval can yield accounts of Native American burning that provide insight into the use and distribution of burning, also yielding a denser spatial distribution of such accounts. To our knowledge, most higher-quality accounts discovered within the western NYS study area have not been cited or listed in major literature on Native American burning. We interpret the patterns and uses of Native American burning in this study as those associated with land-use practices of the late 18th century within the study area. Results reiterated the use of fire for purposes described in previous literature (e.g. Williams 2005), particularly for managing game and facilitating travel; most interestingly this study located an account on the use of fire for exterminating rattlesnakes near Niagara Falls.

The estimated distances from locations of burning to nearest villages (median = 5.6 km, max = 16.2 km) and trails (median = 1.5 km, max = 12.5) were similar to the 5–15 km radius of inferred modifications to vegetation around Native American villages, estimated previously in NYS and Pennsylvania (Black and Abrams 2001; Black et al. 2006; Tulowiecki and Larsen 2015). It must be noted

that the proximity of accounts to Native American trails may be partially due to bias towards early Euro-American travel routes, because these routes developed from routes that Native Americans established. While the number of accounts discovered appears low, this result must also be viewed in context: 12 higher-quality accounts in western NYS alone as compared with Russell's 6 accounts for the northeastern US (1983), Whitney's 33 accounts listed for eastern North America (1996), and Scherjohn et al.'s 87 accounts compiled for North America (2015).

Conclusions

This study demonstrates the potential for information retrieval methods to discover historical accounts of Native American burning within digitized historical documents. Methods are likely adaptable to the search for additional historical accounts on other fire-related topics. This study also shows how such accounts help understand the spatial distribution and purposes of Native American burning. It addresses previous limitations, such as time-consuming reading of documents to discover accounts, lack of indexed databases of accounts, and reliance upon previously-cited accounts. We envision the discovery of historical accounts as useful for enriching other research into the spatiotemporal dynamics of past burning, such as for corroborating records of charcoal in lake sediments (e.g. Munoz and Gajewski 2010) or soil (e.g. Fesenmyer and Christensen 2010), or fire scars in dendrochronological samples (e.g. Brose et al. 2013; Abadir et al. 2019). The discovery of new accounts may also lead to new sites to study using those other methods. Overall, future work should focus on applying information retrieval methods to compile spatial databases of accounts of past Native American burning, with the potential of providing more detailed, localized, and culturally specific understandings of such practices.

Supplementary materials

Files related to this study are available from the authors upon request: a spreadsheet of accounts used to generate lists of key terms; digitized historical documents (.txt format); and bibliographic files for citation management software (i.e. RIS format).

Acknowledgements This research received support from the National Science Foundation under Grant No. 1660388. The authors thank two anonymous reviewers for providing suggestions on an earlier version of this paper.

References

Abadir ER, Marschall JM, Dey DC, Stambaugh MC (2019) Historical fire regimes in red pine forests of the Adirondack Mountains, New York, USA. Nat Area J 39:226–236. https://doi.org/10.3375/043.039.0209

Abrams MD, Nowacki GJ (2008) Native Americans as active and passive promoters of mast and fruit trees in the eastern USA. Holocene 18:1,123–1,137. https://doi.org/10.1177/0959683608095581

Alex B, Grover C, Klein E, Tobin R (2012) Digitised historical text: does it have to be mediOCRe? In: Proceedings of KONVENS 2012 (LThist 2012 workshop). Vienna, Austria, pp 401–409

Black BA, Abrams MD (2001) Influences of Native Americans and surveyor biases on metes and bounds witness-tree distribution. Ecology 82:2,574–2,586. https://doi.org/10.1890/0012-9658(2001)082%5b2574:ionaas%5d2.0.co;2

Black BA, Ruffner CM, Abrams MD (2006) Native American influences on the forest composition of the Allegheny Plateau, Northwest Pennsylvania. Can J For Res 36:1,266–1,275. https://doi.org/10.1139/x06-027

Bowman DMJS, Balch J, Artaxo P et al (2011) The human dimension of fire regimes on Earth. J Biogeogr 38:2,223–2,236. https://doi.org/10.1111/j.1365-2699.2011.02595.x

Brose PH, Dey DC, Guyette RP, Marshall JM, Stambaugh MC (2013) The influences of drought and humans on the fire regimes of northern Pennsylvania, USA. Can J For Res 43:757–767. https://doi.org/10.1139/cjfr-2012-0463

Brown H (2000) Wildland burning by American Indians in Virginia. Fire Manag Today 60:29–39

Cappon LJ (ed) (1976) Atlas of Early American History: The Revolutionary Era, 1760-90. Princeton University Press, Princeton

Clayton WW (1879) History of Steuben County. Lewis, Peck & Co., Philadelphia, New York

Cleveland SC (1873) History and Directory of Yates County, vol 1. S.C Cleveland, Penn Yan, NY

Core Team R (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Cronon W (1983) Changes in the Land: Indians, Colonists, and the Ecology of New England, 1st revised. Hill and Wang, New York

Dalal MK, Zaveri MA (2011) Automatic text classification: a technical review. Int J Comput Appl 28:37–40. https://doi.org/10.5120/3358-4633

Day GM (1953) The Indian as an ecological factor in the northeastern forest. Ecology 34:329–346. https://doi.org/10.2307/1930900

Denevan WM (1992) The pristine myth: the landscape of the Americas in 1492. Ann Assoc Am Geogr 82:369–385. https://doi.org/10.2307/2563351

Doty LR (1905) History of Livingston County, New York: from its earliest traditions to the present, together with Early Town Sketches. W. J Van Deusen, Jackson

Dwight T (1823) Travels in New-England and New-York. W. Baynes and Ogle, London

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Engelbrecht W (2003) Iroquoia: The Development of a Native World. Syracuse University Press, Syracuse

Esri (2017) ArcGIS 10.5.1. Redlands, CA

Feinerer I, Hornik K (2018) tm: Text Mining Package. R package version 0.7-5. https://CRAN.R-project.org/package=tm

Feinerer I, Hornik K, Meyer D (2008) Text mining infrastructure in R. J Stat Softw 25:1–54

Fesenmyer KA, Christensen NL (2010) Reconstructing Holocene fire history in a southern Appalachian forest using soil charcoal. Ecology 91:662–670. https://doi.org/10.1890/09-0230.1

Filby PW (1985) A Bibliography of American County Histories. Genealogical Publishing Company, Baltimore
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1,189–1,232
Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378. https://doi.org/10.1016/s0167-9473(01)00065-2
Grumet RS (1995) Historic Contact: Indian People and Colonists in Today's Northeastern United States in the Sixteenth through Eighteenth Centuries. University of Oklahoma Press, Norman
Harris ZS (1954) Distributional Structure. Word 10:146–162. https://doi.org/10.1080/00437956.1954.11659520
Hays J, Post CF (1999) Journey on the forbidden path: chronicles of a diplomatic mission to the Allegheny country, March-September, 1760. American Philosophical Society, Philadelphia
Hijmans RJ, Phillips S, Leathwick J, Elith J (2013) Dismo: species distribution modeling. R package version 0.8-17
Hill MJ, Hengchen S (2019) Quantifying the impact of dirty OCR on historical text analysis: eighteenth century collections online as a case study. Digital Scholarship in the Humanities fqz024. https://doi.org/10.1093/llc/fqz024
Internet Archive (2019) About the Internet Archive. https://archive.org/about/
Jennings F, Fenton WN (1995) The history and culture of Iroquois diplomacy: an interdisciplinary guide to the treaties of the six nations and their league. Syracuse University Press, Syracuse
Jones EE (2010) An analysis of factors influencing sixteenth and seventeenth century Haudenosaunee (Iroquois) settlement locations. J Anthropol Archaeol 29:1–14. https://doi.org/10.1016/j.jaa.2009.09.002
Keeley JE, Aplet GH, Christensen NL et al (2009) Ecological foundations for fire management in North American Forest and Shrubland Ecosystems. United States Department of Agriculture, United States Forest Service, Pacific Northwest Research Station, Portland
Keister M (1998) Rush Oak openings unit management plan. NYS Department of Environmental Conservation, Bath
Khan A, Baharudin B, Lee LH, Khan K (2010) A review of machine learning algorithms for text-documents classification. J Adv Inf Technol 1:4–20
Lewis HT, Anderson MK (2002) Introduction. In: Lewis HT, Anderson MK (eds) Forgotten fires: Native Americans and the Transient Wilderness. University of Oklahoma Press, Norman
Maude J (1826) Visit to the Falls of Niagara, in 1800. Longman, London
McEwan RW, Dyer JM, Pederson N (2011) Multiple interacting ecosystem drivers: toward an encompassing hypothesis of oak forest dynamics across eastern North America. Ecography 34:244–256. https://doi.org/10.1111/j.1600-0587.2010.06390.x
McIntosh WH (1878) History of Ontario County. Everts, Ensign, & Everts, Philadelphia, PA, New York
McKinstry ER (1997) Personal accounts of events, travels, and everyday life in America: an annotated bibliography. The Henry Francis du Pont Winterthur Museum Inc, Winterthur
Minard JS (1896) Allegany County and its people. A centennial memorial history of Allegany County, New York. W. A. Fergusson & Co., Alfred
Morgan LH (1901) League of the Ho-De'-No-Sau-Nee, or Iroquois. Dodd, Mead and Company, New York
Morton T (1883) New english Canaan. The Prince Society, Boston
Munoz SE, Gajewski K (2010) Distinguishing prehistoric human influence on late-Holocene forests in southern Ontario, Canada. Holocene 20:967–981. https://doi.org/10.1177/0959683610362815
Parker AC (1920) The Archeological History of New York. The University of the State of New York, Albany
Parshall T, Foster DR (2002) Fire on the New England Landscape: regional and temporal variation, cultural and environmental controls. J Biogeogr 29:1,305–1,317. https://doi.org/10.2307/827553
Pyne SJ (1982) Fire in America: a cultural history of wildland and rural fire. Princeton University Press, Princeton
Roberts MF (1891) Historical Gazetteer of Steuben County. New York. Millard F, Roberts, Syracuse, NY
RStudio Team (2016) RStudio: integrated development environment for R. RStudio Inc, Boston
Russell EWB (1983) Indian-set fires in the forests of the Northeastern United States. Ecology 64:78–88. https://doi.org/10.2307/1937331
Scherjon F, Bakels C, MacDonald K, Roebroeks W (2015) Burning the land: an ethnographic study of off-site fire use by current and historically documented foragers and implications for the interpretation of past fire practices in the landscape. Curr Anthropol 56:299–326. https://doi.org/10.1086/681561
Smith BD (2011) General patterns of niche construction and the management of "wild" plant and animal resources by small-scale preindustrial societies. Philos Trans R Soc B-Biol Sci 366:836–848. https://doi.org/10.1098/rstb.2010.0253
Snow DR (1996) The Iroquois. Blackwell Publishers Inc, Malden
Stewart OC (2002) Forgotten fires: Native Americans and the transient wilderness. University of Oklahoma Press, Norman
Tulowiecki SJ (2018) Information retrieval in physical geography: a method to recover geographical information from digitized historical documents. Prog Phys Geogr: Earth Environ 42:369–390. https://doi.org/10.1177/0309133318770972
Tulowiecki SJ, Larsen CPS (2015) Native American impact on past forest composition inferred from species distribution models, Chautauqua County, New York. Ecol Monogr 85:557–581. https://doi.org/10.1890/14-2259.1
Tulowiecki SJ, Robertson DS, Larsen CPS (2019) Oak savannas in western New York State, circa 1795: synthesizing predictive spatial models and historical accounts to understand environmental and Native American influences. Ann Assoc Am Geogr. https://doi.org/10.1080/24694452.2019.1629871
Turner O (1849) Pioneer history of the Holland purchase of western New York: embracing some account of the ancient remains… and a history of pioneer settlement under the auspices of the Holland company; including reminiscences of the war of 1812; the origin, progress and completion of the Erie canal, etc., etc., etc. Jewett, Thomas & Co., Buffalo
Vale TR (2002) The pre-European landscape: pristine or humanized? In: Vale TR (ed) Fire, native peoples, and the natural landscape. Island Press, Washington, DC, pp 1–39
Waite DB (1883) O-Neh-Da Te-Car-Ne-O-Di or Up and Down the Hemlock. G.E. Colvin & G.P, Waite, Canadice, NY
Whitney GG (1996) From coastal wilderness to fruited plain: a history of environmental change in temperate North America, 1500 to the present. Cambridge University Press, New York
Wickham H, François R, Henry L, Müller K (2018) dplyr: A grammar of data manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr
Williams GW (2000) Introduction to aboriginal fire use in North America. Fire Manag Today 60:8–12
Williams GW (2005) References on the American Indian use of fire in ecosystems. USDA Forest Service, Washington, DC
Yale University (2019) eHRAF World Cultures. In: Human Relations Area Files. http://hraf.yale.edu/products/ehraf-world-cultures/. Accessed 11 January 2019