BIBLIOGRAPHIC REFERENCE ANALYSIS IN ARCHIVAL DATA USING SUPERVISED MACHINE LEARNING AND GRAMMATICAL FEATURES

by

James Patrick Philips

December, 2021

Director of Thesis: Nasseh Tabrizi, PhD
Major Department: Computer Science

ABSTRACT

Bibliographic references are integral to scholarly discourse in humanities disciplines. While prior work has focused on reference extraction and parsing, little research has investigated the classification of footnotes containing bibliographic citations and author commentary using supervised machine learning methodologies. For this thesis, we contextualize bibliographic reference analysis within the broader domain of archival document processing through an original literature survey of current techniques, tools, and trends in the field of historical document processing. Next, we review related work on bibliographic citation identification and reference parsing. Finally, using a historiographic dataset drawn from the JSTOR humanities archive, we train and compare the performance of a suite of single and hybrid machine learning classifiers on a novel, previously unexplored bibliographic reference classification task. Moreover, as a part of this analysis, we compare the performance of traditional features and novel, grammatical features drawn from natural language processing. Our work demonstrates the superiority of hybrid models for classification of scholarly footnotes containing historiographic bibliographic references, the transferability of features from reference extraction to this research problem, and the viability of training machine learning models for this task utilizing novel, grammatical features.

BIBLIOGRAPHIC REFERENCE ANALYSIS IN ARCHIVAL DATA USING SUPERVISED MACHINE LEARNING AND GRAMMATICAL FEATURES

A Thesis

Presented to the Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Software Engineering

by

James Patrick Philips

December, 2021

BIBLIOGRAPHIC REFERENCE ANALYSIS IN ARCHIVAL DATA USING SUPERVISED MACHINE LEARNING AND GRAMMATICAL FEATURES

by

James Patrick Philips

APPROVED BY:

DIRECTOR OF THESIS _____
Nasseh Tabrizi, PhD

COMMITTEE MEMBER _____
Qin Ding, PhD

COMMITTEE MEMBER _____
Venkat N. Gudivada, PhD

COMMITTEE MEMBER _____
Mark Hills, PhD

CHAIR OF THE DEPARTMENT OF COMPUTER SCIENCE _____
Venkat N. Gudivada, PhD

DEAN OF THE GRADUATE SCHOOL _____
Paul J. Gemperline, PhD

To my parents,

Frank Parker Philips, III,

and

Carolyn Staton Philips,

with deepest love and gratitude

ACKNOWLEDGEMENTS

This thesis evolved out of my valuable experience in the NSF Research Experience for Undergraduates in Software Testing and Analytics held by East Carolina University's Computer Science Department during the Summer of 2018. I am deeply grateful to Dr. Nasseh Tabrizi for his ongoing mentorship, kindness, encouragement, and sage advice throughout my academic journey and the drafting of this thesis. His enthusiasm for research and the breadth of his research interests are an inspiration. Likewise, I appreciate Dr. Qin Ding, Dr. Venkat Gudivada, and Dr. Mark Hills's willingness to serve on my defense committee. Each of them has been an important influence in my studies of computing and software engineering. Moreover, I wish to acknowledge with appreciation the influence of Dr. Ronnie Smith, whose Natural Language Processing elective in Spring 2018 had a formative impact on my interests in the use of software to analyze human language.

Many family members and friends have offered their encouragement throughout my graduate studies and the work on this thesis. Likewise, I have benefitted from the encouragement and friendship of several of Dr. Tabrizi's current and former students in his research group, including: Storm Davis, Nate Fenwick, Sara Jooneghani, Babak Malekishoja, and Maryam Navaei. Moreover, my friends, Derek Nelms and Joseph Paramore, have provided ongoing camaraderie and encouragement throughout my studies. Several professors from ECU's Department of History have offered timely encouragement as well, including Dr. Jonathan Reid, Dr. Frank Romer, and Dr. Michael Bennett. From Edgecombe Community College, my former instructors Trey Cherry, Monika Fleming, Stephen Herring, and Rebecca Stamilio-Ehret have been continuously supportive throughout my academic journey in Tarboro and Greenville alike. Dr. Doug Proffit has had an abiding influence on my studies since my time as a student at Falls

Road Baptist Church School. His rigorous instruction in the craft of research and his personal example of Christian scholarship has indelibly shaped this work and many projects before it for which I remain thankful.

Finally, this thesis is dedicated to my parents, Frank Parker Philips, III, and Carolyn Staton Philips. Their love, support, and encouragement throughout my academic endeavors have been unceasing, and I am deeply grateful to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

The advent of the "Big Data" epoch has transformed computing. Non-relational databases have become popular alternatives to traditional, relational ones to accommodate the need to scale data stores [Gudivada et al 2015]. It has driven an explosive growth in the use of machine learning and deep learning to leverage the vast quantities of data available across many scientific and commercial domains [Gudivada et al 2016]. It has necessitated new microservices architectures in software engineering since traditional monolithic architectures cannot adequately scale [Jamshidi et al 2018].Developing methodologies to mine this rich reservoir of data at a massive scale, developing robust system architectures to retrieve information from it efficiently, and then analyze this information ethically and insightfully are key imperatives of computer scientists, software engineers, and data scientists in the twenty-first century.

Yet, many computing practitioners and researchers have a myopic focus on contemporary Big Data. They are oblivious to the Big Data of the past that antedates to the era of modern computing inaugurated by Alan Turing, John von Neumann, Thomas Kilburn, and other pioneering scholars [Ceruzzi 2003; Anderson 2009]. Traditionally the domain of historians and other researchers in the humanist disciplines, this archival data is "forgotten data" that awaits its own Big Data renaissance. Unlocking its potential will necessitate collaboration between researchers in computing and the humanities [Terras et al 2018]

1.1 Conceptualizing Archival Data as "Big Data"

Throughout history religious and secular libraries have been integral to the preservation of human knowledge and the epicenters of renewals of learning. For example, monastic *scriptoria* in Europe were crucial preservers of Greek and Roman literature following the collapse of the western Roman Empire in the fifth century AD [Scrivner 1980]. Monastic scribes

produced handwritten manuscripts that kept knowledge of antiquity alive, and renewed focus on these manuscripts helped to spark the fifteenth century Renaissance. Medieval manuscripts were supplemented by the explosive number of new published works that flowed off the printing presses of Europe during and after the Renaissance. Libraries became archival repositories for all kinds of human learning, much of which continues to lie dormant and inert, awaiting scholarly attention. Just like contemporary Big Data, archival data exists in unstructured, semi-structured, and structured varieties. This archival data in all its forms is the "Big Data" of the past. Conceptualizing it requires transcending the chronological shortsightedness that is an affliction of our modern moment.

1.2 Historical Documents as Archival Data

Historical documents represent one embodiment of archival data. Medieval and Enlightenment-era manuscripts as well as printed texts have drawn the interest of computing scholars interested in handwriting recognition, document layout analysis, and optical character recognition. A more extensive treatment of this field is given in Chapter 2 of this thesis. However, much of the research in this domain seems fixated on novel algorithms to improve a particular phase of historical document analysis and less concerned with the analysis of the semantic content of this archival data. To ignore the meaning embodied in these historical documents is to cut ourselves off from vital insights they have to share.

1.3 Bibliographic References as Scholarly Archival Data

Not all archival data remains cryptic. The advent of digital libraries has enabled archival data from the more recent past to be stored in databases and accessed within information retrieval systems. Digital libraries are integral to the workflow of scholarship in both the sciences and the humanities, yet the revolutionary effects of Big Data are felt here as well. Researchers in

Bioinformatics have been in the vanguard of developing novel methodologies in text mining to extract useful information from these vast digital libraries of scholarly archival data [Xie et al 2013]. Bibliographic references are one information type that can be mined from these semi-structured scholarly archival data. Chapter 3 of this thesis examines related work on methodologies for extraction and parsing of these references in both the bioinformatics and increasingly humanities domains, especially that of historiography.

1.4 Thesis Motivation

This thesis was motivated by the author's conviction that archival data matters in contemporary discourse. As medieval literature scholar C.S. Lewis observed, old books, and therefore, by extension data from the past, are vital in the quest for truth: "Every age has its own outlook. It is especially good at seeing certain truths and specially liable to make certain mistakes. We all, therefore, need the books that will correct the characteristic mistakes of our own period. And that means the old books." [Lewis 1970] Developing methodologies and software systems to support historians and other scholars in humanities disciplines will be essential to unlocking the secrets of this "forgotten data." Yet, this is a joint effort in "archival analytics" that requires the algorithmic expertise of computer scientists, the systems architecture expertise of software engineers, and the data analysis insight of data scientists to complement the domain knowledge of historians and other humanists.

The computing field has antecedents for this kind of interdisciplinary, collaborative work in its early history. For example, Fr. Roberto Busa collaborated with IBM during the 1950s and beyond to use their mainframes to create a vocabulary index to the entire 10,600,000-word Latin corpus of medieval theologian Thomas Aquinas [Busa 1980; Jones 2018]. Moreover, biblical scholar John Ellison used the Mark 1 at Harvard University to analyze the transmission of

manuscript readings within the Christian New Testament [Bowles 1967]. Humanities computing is not new. A new generation of computing practitioners and software engineers simply need to rediscover and utilize it, creating novel methodologies and building software systems to recover the "forgotten data" in the archives of the past.

1.5 Research Questions and Thesis Contributions

Despite the excellent work in text mining to extract bibliographic references from archival documents and work in machine learning to parse the metadata of these citations, there are several facets of bibliographic references that have yet to be explored. This thesis seeks to examine the following research questions:

- RQ1: Can we predict the type of bibliographic footnote based on the properties of its citations and semantic content?
- RQ2: Can features traditionally used for bibliographic reference extraction be transferred to this footnote classification task?
- RQ3: Can novel, grammatical features drawn from the field of Natural Language Processing (NLP) be utilized to predict the type of bibliographic footnote?

Chapter 4 discusses the methodology for addressing these questions, including the creation of a novel dataset of archival historiographic bibliographic references. Chapter 5 discusses the results of the experimental phase of the research for this thesis. This thesis makes the following original contributions:

- Contribution 1: We train a suite of single and hybrid supervised machine learning classifiers and evaluate their performance on a novel bibliographic reference classification task. These experiments demonstrate the superiority of hybrid machine learning models for this kind of bibliographic reference analysis.

- Contribution 2: We demonstrate that features used for bibliographic reference extraction are transferable to the differentiation of reference string type.
- Contribution 3: We demonstrate that novel, grammatical features are viable alternatives to traditional feature sets for training supervised machine learning classifiers.

1.6 Thesis Structure

This thesis is structured as follows. Chapter 2 explores computational archival data analysis through a literature survey of techniques, tools, and trends utilized in historical document processing. Chapter 3 discusses related literature on bibliographic reference extraction and parsing techniques from archival data. Chapter 4 reviews the methodology used for the experimental phase of this thesis. Chapter 5 discusses the results of the thesis experiment. Finally, Chapter 6 concludes with a summary of the work and suggests directions for future work.

Chapter 2: A Survey of Techniques, Tools, and Trends in Historical Document Processing

2.1 Introduction

Historical Document Processing (HDP) is the process of digitizing written and printed material from the past for future use by historians. Digitizing historical documents preserves them by ensuring a digital version will persist even if the original document is destroyed or damaged. Since many historical documents reside in libraries and archives, access to them is often hindered. Digitization of these historical documents thus expands scholars' access to archival collections as the images are published online and even allows them to engage these texts in new ways through digital interfaces [Chandna et al 2016; Tabrizi 2008]. HDP incorporates algorithms and software tools from various subfields of computer science to convert images of ancient manuscripts and early printed texts into a digital format usable in data mining and information retrieval systems. Drawing on techniques and tools from computer vision, document analysis and recognition, natural language processing, and machine learning, HDP is a hybrid field. This chapter surveys the major phases of HDP, discussing techniques, tools, and trends. After an explanation of our research methodology, digitization challenges, techniques, standard algorithms, tools, and datasets are discussed, and concludes with suggestions for further research. This chapter was previously published in the proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) [Philips and Tabrizi 2020].

2.2 Methodology

2.2.1 Research Rationale

This chapter examines the evolution of the techniques, tools, and trends within the HDP field over the past twenty-two years (1998-2020). The author believes this extended scope is

warranted: No prior study was found that summarized the HDP workflow for both handwritten archival documents and printed texts. Prior studies have focused on one dimension of the problem, such as layout analysis, image binarization, or actual transcription. Very few discussed aspects of a full historical document processing workflow.

2.2.2 Article Selection Criteria

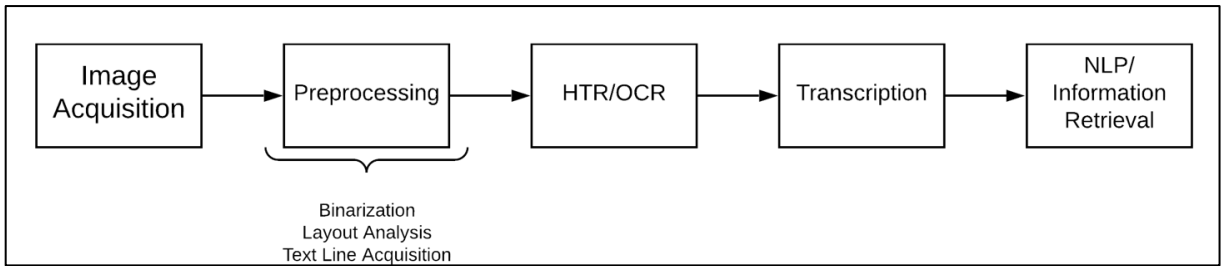
This research focuses on historical documents written in Latin, medieval and early modern European vernaculars, and English reflecting the current state of the HDP field: most of the work on historical archival documents has focused on western scripts and manuscripts. From the initial collection of 300+ articles chosen, 50 were selected for this survey. This survey emphasizes the computer science dimension of HDP, especially machine learning methodologies, software tools, and research datasets. The authors envision other computer scientists, digital humanists, and software developers interested in HDP and cultural heritage as their primary audience.

2.3 Techniques and Tools

2.3.1 Archival Document Types and Digitization Challenges

Historical documents broadly defined include any handwritten or mechanically produced document from the human past. Many have been preserved in the archives of museums and libraries, which have pursued extensive digitization efforts to preserve these invaluable cultural heritage artifacts. An enduring goal within the field of document image analysis has been achieving highly accurate tools for automatic layout analysis and transcription [Baechler and Ingold 2010].

Figure 2-1: The steps in a conventional HDP workflow for handwritten and printed documents.



A typical HDP workflow proceeds through several sequential phases. After image acquisition, the document image is pre-processed and handwritten text recognition (HTR) or optical character recognition (OCR) is performed. This phase yields a transcription of the document's text. This transcription is the input to natural language processing and information retrieval tasks.

Prior to the 15th century, the majority of historical documents were texts produced by hand. After Gutenberg's printing press, published works were produced on the printing presses while private documents continued to be done by hand. This dichotomy in document types beginning in the Early Modern era led to diverse document types that must be dealt with differently during the HDP process.

The eclectic nature of all handwritten documents challenges automatic software tools. Medieval manuscripts are often more legible, and the inter- character segmentation of minuscule script are easier to train machine learning-based classifiers for than the continuous cursive of early modern handwritten texts. However, significant challenges in medieval documents are their complex layouts and intricate artwork [Simistira et al 2016]. Continuous cursive script in Early Modern documents is challenging during the HTR phase, while medieval documents present greater challenges during layout analysis. Other challenges with historical documents include

bleed-through from the opposite sides of the pages, illegible handwriting, and image resolution quality.

The earliest printed texts, known as incunabula, have posed the most difficulties for accurate, digital transcription of printed works [Rydberg-Cox 2009]. Their fonts differ vastly from modern typefaces, and modern OCR software produces poor recognition results. The extensive use of textual ligatures also poses difficulties since they declined in use as printing standardized. After 1500 greater uniformity came to printed books, and by the early 19th century, the mass production of printed texts led to books that modern layout analysis and OCR tools could reliably and consistently digitize at scale, as seen in the digitation efforts of the Internet Archive and Google Books in partnership with libraries [Bamman and Smith 2012]. This opens up possibilities for Information Retrieval in archival "Big Data."

2.3.2 Techniques

2.3.2.1 Pre-processing Phase

This pre-processing phase normally includes binarization/thresholding applied to the document image, adjustment for skew, layout analysis and text- line segmentation. Various studies have proposed various binarization methods including Bolan et al 2010, Messaoud et al 2012, and Roe and Mello 2013. Dewarping and skew reduction methods have been proposed in studies including Bukhari et al 2011 and performance analysis conducted in Rahnemoonfar and Plale 2013. Layout analysis is one of the most challenging aspects of HDP. Recent work has also examined the use of neural networks to restore degraded historical documents [Raha & Chanda 2019]. Due to their complex page layouts, many studies have focused on layout analysis tools, algorithms, and benchmark datasets especially for medieval documents. Baechler and Ingold proposed a layout model for medieval documents. Using manuscript images from the E-codices

project, they modeled a medieval manuscript page as several "layers": document text, marginal comments, degradation, and decoration. Overlapping polygonal boxes are used to identify the constituent layers and are represented in software via XML.

Gatos et al 2014 developed a layout analysis and line segmentation software module designed to produce input to HTR tools. Their work was incorporated into the Transcriptorium project's Transkribus software. Pintus, Rushmeier, and Yang likewise explore layout analysis and text-line extraction with an emphasis on medieval manuscripts. Pintus et al 2015 address the problem of initial calculation of text- line height. They segment the text regions coarsely and apply a SVM classifier to produce a refined text line identification. They note their method is not adversely affected by skewed texts and usually does not necessitate any alignment correction.

Yang et al (2017) extend their work on text-height estimation and layout analysis to an automated system that can work on a per-page basis rather than per manuscript. They propose three algorithms, one for text-line extraction, one for text block extraction, and one for identifying "special components." These use semi-supervised machine learning technique and focus on medieval manuscripts produced originally by professional scribes. Their results demonstrate that the desideratum of automatic algorithmically-layout analysis with high precision, recall, and accuracy is drawing nearer to reality.

2.3.2.2 Handwritten Text Recognition

Due to the inherent challenges of HTR for historical documents, some studies including [Rath and Manmatha 2006; Fischer et al. 2012] explored keyword spotting techniques as an alternative to producing a complete transcription. Early keyword spotting techniques approached it as an image similarity problem. Clusters of word images are created and compared for similarity using pairwise distance. Fischer et al explored several data-driven techniques for both

keyword spotting and complete transcription [Fischer et al. 2009, 2012, 2014]. One problem with word-based template matching is that the system can only recognize a word for which it has a reference image. Rare (out of vocabulary) words cannot be recognized. As a solution, the HisDoc project applied character-based recognition with Hidden Markov Model (HMM) to keyword spotting. For their keyword spotting analysis, they compared the character-based system with a baseline Dynamic Time Warp (DTW) system. Using Mean Average Precision as their evaluation metric, they found that the HMMs outperformed the DTW system on both localized and global thresholds for the George Washington and Parzival datasets (GW: 79.28/62.08% vs 54.08/43.95% and Parzival 88.15/85.53% vs 36.85/39.22%). The HisDoc project also compared HMMs and neural network performance on the University of Bern's Historical Document Database (IAM-HistDB) to produce full transcriptions. They used a Bi-directional Long Short-term Memory (BLSTM) architecture that could mitigate the vanishing gradient problem of other neural network designs. Each of their nine geometric features used for training corresponds to an individual node in the input layer of the network. Output nodes in the network correspond to the individual characters in the character set. The probability of a word is computed based on the character-probabilities. According to [Fischer, Naji 2014], word error rates were significantly better for the neural network architecture than the HMM system on all three sets of historical document images: St. Gall 6.2% vs 10.6%, Parzival 6.7% vs 15.5%, and George Washington 18.1% vs 24.1%.

Neural networks continue to be the ascendant technique within the field for HTR. Granell et al. 2018 examined the use of convolutional recurrent neural networks for late medieval documents. The convolutional layers perform automatic feature extraction which precludes the need for handcrafted geometric or graph-based features such as those used by HisDoc. For deep

neural network architectures to be competitive for time efficiency with other techniques, they require significant computational power. This is obtained through the use of a GPU rather than a CPU. Working with the Rodrigo dataset, they achieved their best results using a convolutional neural network supplemented with a 10-gram character language-model. Their word error rate was 14%.

2.3.2.3 Historical Optical Recognition

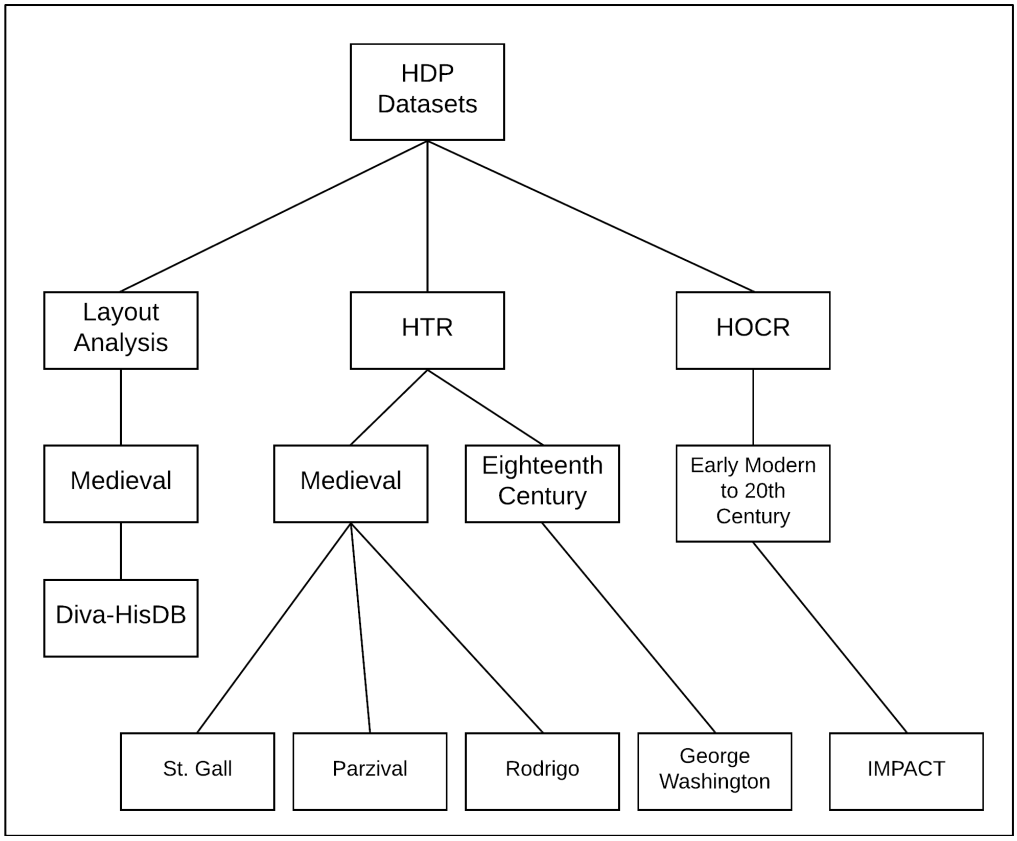
As with HTR, historical OCR can be accomplished with several techniques. However, neural network- based methods have become more prominent in the software libraries and literature recently. Since printed texts in western languages rarely use scripts with interconnected letters, segmentation-based approaches are feasible with OCR that are not practical for HTR. Nevertheless, historical OCR is drastically more difficult than modern OCR [Springmann and Lüdeling 2017]. One challenge is the vast variability of early typography. Historical printings not laid out with modern, digital precision, and a plethora of early fonts were utilized across Europe [Christy et al 2017]. A multitude of typeface families exist, including Gothic script, Antiqua, and Fraktur. Although printing techniques standardized in the early 19th century, printed documents from 15th- 19th centuries are too idiosyncratic for OCR machine learning classifiers trained using modern, digital fonts. Among the most difficult historical texts for OCR are incunabula due to their extensive use of ligatures, typographical abbreviations derived from medieval manuscripts that do not always have a corresponding equivalent in Unicode, and unpredictable word-hyphenation across lines [Rydberg- Cox 2009]. The model training limitations of commercial software such as Abbey Fine Reader mean that researchers must resort to open source alternatives such as Tesseract or OCRopus [Springmann et al. 2014]. Tesseract's classifier can be trained using either synthetic data (digital fonts that resemble historical ones) or

with images of character glyphs cropped from actual historical text images. Tesseract and OCRopus both offer neural network classifiers Although high accuracies are achievable with neural networks, some of the same caveats apply from their use for HTR. These classifiers require substantial training data, with the corollary of extensive ground truth that must be created manually, and this classifier is computationally intensive for CPUs [Springmann et al 2014].

2.3.2.4 Software Tools and Datasets

Several software tools and datasets (Figure 2-2) exist for researchers and practitioners pursuing historical document processing.

Figure 2-2: A taxonomy of HDP datasets based on use case and time-period.



For historical OCR, these include Abbey FineReader, Tesseract, OCRopus, and AnyOCR tools and primarily the IMPACT dataset of early modern European printed texts. Few generic tools exist for historical HTR tasks, but researchers do have access to the IAM-HistDB and Rodrigo

datasets. These variously contain images of full manuscript pages, individual words and characters, and corresponding ground truth for medieval Latin and early German and Spanish manuscripts. The IAM-HistDB also contains the Washington dataset for historical cursive handwriting recognition. In addition to software and datasets for the transcription phase of historical document processing, the Alethia tool and the IMPACT and Diva-HistDB datasets can be used for researching layout analysis and other pre- processing tasks. The rest of this section surveys the characteristics of the available datasets and discusses training, testing, and evaluation methodologies.

Few options exist for researchers seeking to work with medieval manuscript transcription. Two medieval datasets are included in the IAM-HistDB. The St. Gall dataset features images of a ninth century Latin manuscript written in Carolingian script by a single scribe. Fischer et al utilized the images and corresponding previously published page transcriptions from J.P. Migne's *Patrologia Latina* to create the dataset [Fischer et al. 2011]. In addition to page images and transcription, the dataset includes extensive ground-truth: text-lines and individual word images have been binarized, normalized, and annotated with line-level transcription. Originally developed by the HisDoc project, the dataset has since been used in further research.

While Latin was the dominant ecclesiastical and scholarly language of Europe during the medieval period, some literature was produced in the vernacular languages. Two datasets exist for researchers investigating HTR in those vernacular texts, specifically the Old German and Old Spanish dialects. Included with the IAM-HistDB, the Parzival dataset contains manuscript pages of an Arthurian epic poem written in Old German from the 13th and 15th centuries. The 47 Parzival images are drawn from three different manuscripts produced by three scribes using

Gothic minuscule script in multi- column layouts. Like the St. Gall set, the Parzival collection includes page images and transcription along with ground truth annotation. Text-lines and single word images have been binarized, normalized, and annotated with a full line-level transcription. Known as the Rodrigo corpus, the Old Spanish dataset is larger than either the St. Gall or Parzival datasets at 853 pages. Created for HTR and line extraction research, the researchers based at the Universitat Politecnica de Valencia used the digitized images of an Old Spanish historical chronicle, the "Historia de Espanana el archbispo Don Rodrigo" [Serrano et al 2011]. The manuscript is from 1545, and thus can be traced to the emergence of printing press technology. Although the creators of the dataset published results of running a hybrid HMM-based image classifier with a language model, Granell et al have used the dataset with deep neural networks [Granell et al 2018].

The Washington dataset is the third dataset included in the IAM HistDB. Drawn from the George Washington papers at the US Library of Congress, its script is continuous cursive in the English language. First used in Rath and Manmatha, the HistDoc project supplemented the dataset with individual word and text-line images and corresponding ground truth transcriptions for each line and word [Fischer et al 2010]. The Washington dataset is especially valuable for cursive HTR in historical documents.

The previously described IAM-HistDB datasets dealt exclusively with historical HTR. As a benchmark for evaluating pre-processing performance on medieval documents, the HistDoc project created the Diva-HistDB. This dataset contains 150-page images from three different manuscripts with accompanying ground truth for binarization, layout analysis, and line segmentation [Simistira et al 2016]. Written in Carolingian script, two of the manuscripts are from the 11th century, and one from the 14th century written in Chancery script. All three

manuscripts have a single column of text surrounded by extensive marginal annotation. Some pages have decorative initial characters. The layouts are highly complex. The ground truth concentrates on identifying spatial and color-based features. Like the IMPACT dataset, the ground truth is encoded in the PAGE XML format. The dataset is freely available on the HistDoc project website.

While most of the HTR and OCR datasets discussed in this section have focused on Latin languages or Latin script, a dataset has been created for HTR and OCR of historical polytonic (i.e. multiple accents) Greek texts. Introduced by Gatos et al, the dataset was developed for research on the word and character recognition as well as line and word segmentation (Gatos et al 2015). It features 399 pages of both handwritten and printed Greek text, mostly from the nineteenth and twentieth century.

2.3.2.5 Methodologies for Evaluation

Several metrics are used to evaluate the performance of a historical document processing system. For handwritten text recognition systems that use image similarity, precision and recall are two important performance measures. Precision ascertains how many of all the relevant results in the dataset were actually retrieved. For machine learning systems, transcription performance is evaluated using the character error rate, word error rate, or sometimes both if a language model is utilized to enhance the recognition results. Layout analysis performance is assessed using the line error rate and segmentation error rate [Bosch et al 2014].

2.3.2.6 Software Systems

Cultural heritage practitioners seeking production- ready tools for their own historical document preservation projects have two software systems available that provide a full suite of tools for pre- processing, machine learning training, and transcription. These two tools are

DIVA-Services [Würsch et al 2017] and the Transkribus platform from the EU-sponsored READ project [Kahle et al 2017].

DIVA-Services and Transkribus offer similar feature sets to the cultural heritage community. However, they should not be seen as direct competitors. As a cross-platform software service, Transkribus is likely the better solution for archivists seeking an integrated HDP toolchain that requires minimal or no custom software to be developed. Since it offers multiple tools for each step in the HDP process and supports standard formats such as PAGE, it is ideally suited for archivists who need a reliable service for a historical document transcription project that allows support for machine learning training on new datasets. Due to the platform's hybrid open source-closed source nature and lack of tool modularity (users cannot substitute their own libraries directly for a Transkribus one), users who need more flexibility and alignment with open source values may find DIVA-Services more suited to their needs. Since DIVA-Services provides separate API calls for each discrete step in the HDP workflow, this service is more suitable for computer science researchers and archivists who need to integrate existing methods alongside custom software. DIVA-SERVICES and Transkribus thus offer complementary approaches that meet the different use cases of members of the cultural heritage community.

2.4 Recent Trends

Within the past decade, several research projects have advanced the field of historical document processing through the creation of datasets, the exploration of improved techniques, and the application of existing tools to digital archival document preservation efforts. The HisDoc family of projects have made significant contributions to algorithms, tools, and datasets for medieval manuscripts. The inaugural HisDoc project lasted from 2009 to 2013 and

concurrently studied three phases of HDP: layout analysis, HTR, and document indexing and information retrieval [Fischer Nijay et al 2014]. While much of their research focused on medieval documents and scripts, their goal was to create "generic methods for historical manuscript processing that can principally be applied to any script and language." (83)

HisDoc 2.0 was conceived as a direct extension of the original HisDoc project. Concentrated at the University of Fribourg, the focus of this project was advancing digital paleography for archival documents [Garz et al 2015]. The HisDoc 2.0 researchers recognized that historical manuscripts are complex creations and require multi-faceted solutions from computer science. Written by multiple scribes and due to inconsistent layouts, many documents do not conform to the ideal characteristics explored during the first HisDoc project. With HisDoc 2.0, the researchers investigated combining text localization, script discrimination, and scribal recognition into a unified system that could be utilized on historical documents of varying genres and time periods. The HisDoc 2.0 project made several contributions to the field. One was DivaServices, a web service offering historical document processing algorithms with a RESTful (representational state transfer) API to circumvent the problem many developers and practitioners face with the installation of complicated software tools, libraries, and dependencies [Würsch et al 2016]. Another contribution was the DivaDesk digital workspace, GUI-based software that makes computer science algorithms for ground truth creation, layout analysis, and other common tasks accessible for humanities scholars [Eichenberger et al 2014]. The project explored ground truth creation, text region and layout analysis with neural networks, and aspects of scribal identification. Finally, the project produced and released the Diva-HisDB dataset.

The IMPACT project was a European Union- funded initiative to develop expertise and infrastructure for libraries digitizing the textual heritage of Europe. Despite the rapid rate of text

digitization by European libraries, the availability of full-text transcriptions was not keeping pace. With many libraries solving the same digitization challenges, solutions to problems were being duplicated, leading to inefficient use of time and resources. Moreover, existing OCR software produced unsatisfactory accuracy for historical printed books. Through the formation of a pan- European consortium of libraries, the IMPACT project consolidated digitization expertise and developed tools, resources, and best practices to surmount the challenges of digitization on such an extensive scale. The project lasted from 2008- 2012. Among its achievements were the monumental creation of the IMPACT dataset of historical document images with ground truth for text and layout analysis, the development of software tools for layout analysis, ground truth creation, and optical character recognition post-correction, the proposal of the PAGE format, and the exploration of techniques for OCR, layout analysis, and image correction [Papadopoulos 2013; Pletschacher & Antonacopoulos 2010; Vobl et al 2014].

The Early Modern OCR Project (eMOP) was an effort by researchers at Texas A & M University to produce transcriptions of the Early English Books Online and 18th Century Collections Online databases. Containing nearly 45 million pages collectively, these two commercial databases are essential tools for historians studying the literature of the 15th through the 18th century. The project produced accurate transcriptions paired with the corresponding text images and made available for crowd-sourced post-correction on the 18thConnect website using the TypeWright tool; it developed a true "Big Data" infrastructure to take advantage of high-performance computing resources for both OCR and image post-processing. Another important contribution was the pioneering work on a historical font database [Heil and Samuelson 2013].

Historical Document Processing transforms scanned documents from the past into digital transcriptions for the future. After pre-processing through binarization, layout analysis, and line

segmentation, the images of individual lines are converted into digital text through either HTR or OCR. Within the past decade, first conventional machine learning techniques using handcrafted features and more recently neural network-driven methodologies have become solutions to producing accurate transcriptions from historical texts from medieval manuscripts and fifteenth-century incunabula through early modern printed works. Projects such as IMPACT, Transcriptorium, eMOP, and HisDoc have made significant contributions to advancing the scholarship of the field and creating vital datasets and software tools. The combined expertise of computer scientists, digital humanists, historians, and archivists will all be necessary to meet the challenge of HDP for the future.

As archives continue to be digitized, the volume and variety of archival data and the velocity of its creation clearly indicate that this is a "Big Data" challenge. Accurate transcriptions are a prerequisite for meaningful information retrieval in archival documents. The creation of robust tools and infrastructure for this new phase of historical document processing will be the mandate of all those who wish to preserve humanity's historical textual heritage in the digital age. The next chapter will examine methodologies for analysis of archival data (specifically bibliographic references) once this data has been created using an HDP workflow

Chapter 3: Related Literature on Bibliographic References

Bibliographic references are integral to research and scholarly discourse in both the sciences and the humanities. Through them, researchers acknowledge their intellectual debts and orient their readers to the extant research literature foundational to their own projects. The citation of other authors' works enables researchers to contextualize their work in the ongoing research discourse of their disciplines and to differentiate it from what has been done before them. These references furnish a mechanism for authors to acknowledge the seminal works of prior scholarship and signal those works that represent novel research directions. Taken together, a work's bibliographic citations create a mosaic, a snapshot of the state of a discipline at a particular point of time.

The preceding chapter examined the general process for digitizing archival documents. This chapter focuses on methodologies for archival data analysis of bibliographic references. Specifically, it defines bibliographic references, contrasts their different uses in scientific and humanities research, and discusses related literature on the extraction and parsing of these references in humanities disciplines from archival data.

3.1 Defining Bibliographic References and Scholarly Citations

Fundamentally, a bibliographic reference is a scholarly citation of another research work. Its key elements include the cited work's author(s), its title, publication details, page references, date of publication, and in the case of digital sources, a DOI. Depending on the research discipline, bibliographic references appear as parenthetical citations or note numbers linked to a list of works cited or list of references appended to the publication's end or contained in footnotes at the bottom of each page.

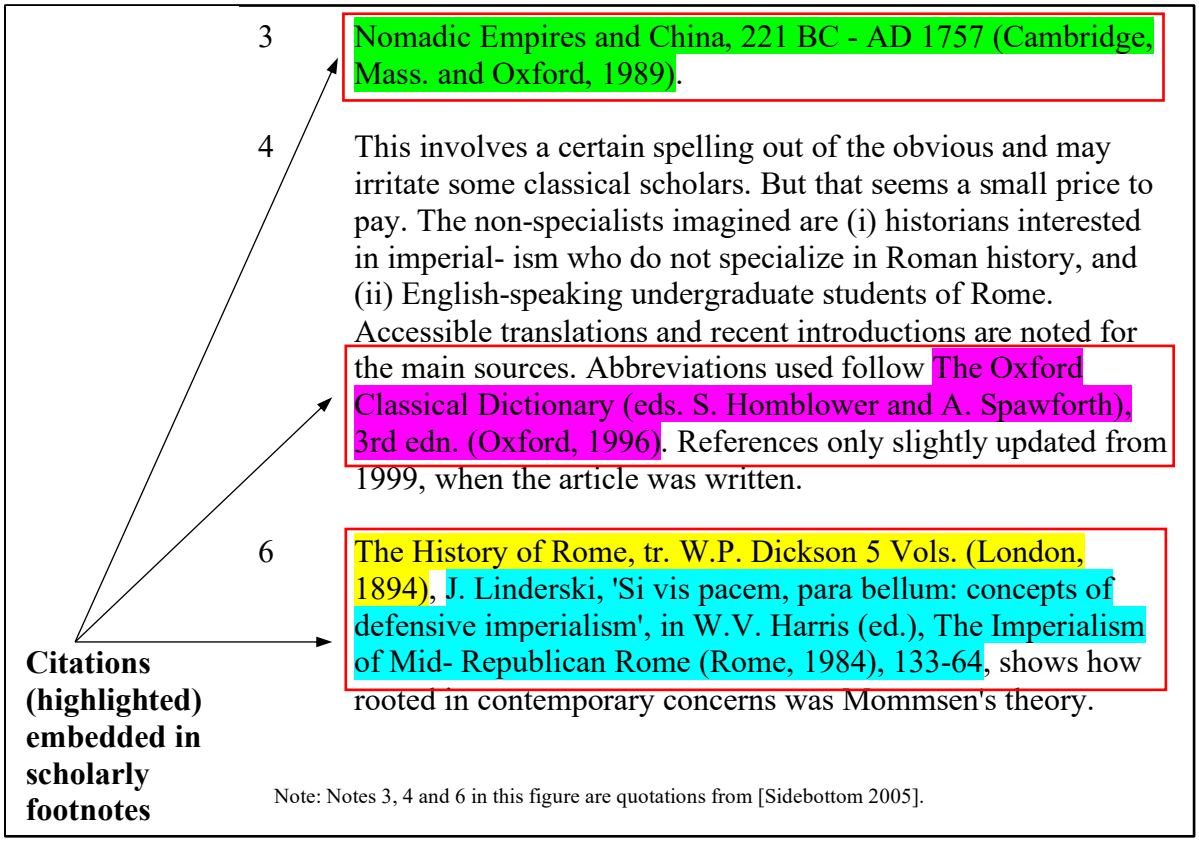
3.2 Role of Bibliographic References in Humanities Scholarship

Within scientific fields, bibliographic references primarily fulfill the role of citation. Like pointers in the venerable C programming language, bibliographic references acting as citations direct a reader to related research or resources (i.e. datasets) in the "heap" of academic memory. Citations permit an author to establish the parameters of their own research through referral to others' work, drawing upon critical ideas, insights, and results without having to recapitulate the entirety of those earlier studies. Since bibliographic references are utilized primarily in the mode of citation in scientific research, they are usually cited minimally in the text of a journal article through a parenthetical reference or note number with the references themselves consigned to a section of endnotes at the work's end.

In contrast, within many humanities disciplines such as history and literary studies, bibliographic references have a much more extensive role than they do in the sciences. While they fulfill the role of citation as do their scientific counterparts, bibliographic references likewise are often embedded within and accompanied by extensive author commentary. This discursive analysis enhances the arguments and discussion of the article or monograph's body. Citations and commentary meld together in humanistic discourse. They link an author to the ongoing scholarly conversation of the *respublica litterarum* that extends throughout time and space [Edelstein et al 2017; Grafton 2009]. Therefore, bibliographic references in humanities disciplines are frequently chronologically diverse. They can contain a plethora of citations to sources drawn from ancient, medieval and Renaissance, and modern eras. Usually located in footnotes or endnotes, these augmented bibliographic references complement and supplement the primary narrative and expository elements of the monograph or research article with crucial semantic content. While multiple citations occur in the sciences, humanistic scholarship

particularly in history and philology frequently embed a plethora of citations and commentary within an extensive matrix of footnotes. Figure 3-1 shows examples taken from an article in the field of Classical Studies on the Roman imperial expansion [Sidebottom 2005]. The footnote text extracted from the article and depicted in the figure show several note types, including single citation (note 3), single citation with scholarly commentary (note 4), and a note with multiple citations and commentary (note 6).

Figure 3-1



As this figure demonstrates, humanities scholarship utilizes footnotes to embed a rich research discourse in parallel with the main text of the publication. Instrumental as a means of citation, bibliographic references in footnotes furnish a vital venue for the research conversation of the humanities disciplines. Some researchers have begun investigating how these citations can be

utilized in information retrieval to build scholarly literature recommender systems [Ollagnier 2016; Ollagnier 2018].

3.3 Methodologies for Reference Extraction and Parsing

Cognizant of the vital importance of bibliographic references and their rich semantic content, researchers in both computer science and bioinformatics have developed a rich literature investigating how to extract bibliographic references from published works and how to parse those extracted references into their constitutive elements. Originally, many of these endeavors were motivated by mining citations for bibliometric analysis [McBurney and Novak 2002]. Bibliographic references are thus envisioned as nodes in a research graph. However, recent work by some digital classicists and other humanities scholars has sought to elaborate and apply these methodologies to bibliographic references [Romanello 2013]. This section highlights methodologies and datasets for reference extraction and parsing.

3.3.1 Reference Extraction

Within the context of a digital library archival system, the semantic content of bibliographic references furnishes a rich trove of data that can be mined and utilized for information retrieval and citation recommendation. Once publications such as journal articles and monographs have been digitized, the citations must first be extracted from the publication. Once extracted, each citation must be parsed into its individual elements. Only then can this archival, bibliographic data become useful information in digital systems.

Reference extraction methodologies in the humanities must account for complex document structures [Gupta et al 2009]. Parenthetical and indirect citations in the body of a work must be linked to the full bibliographic entries. Many footnotes contain multiple citations that are nested within author commentary. Approaches to extraction must reckon with not just extracting

the reference list from the end of a scholarly document as in scientific publications. Footnotes must be isolated from the body text of a page and from each other. This task can be exacerbated for historical scholarly documents by poor page scans and inaccurate optical character recognition (cf. Chapter 2). Approaches have included page layout analysis of PDFs as well as analysis of the extracted full-text of documents [Tkaczyk 2015; Lopez 2009]. Romanello applied conditional random fields (CRFs) to reference extraction in Classics scholarly articles [Romanello et al 2009].

3.3.2 Reference Parsing

Parsing bibliographic references in the humanities is a much more challenging task than in the sciences. While citations share common elements, no universal schema for reference formatting exists among humanities scholars. Furthermore, reference styles have varied significantly over time. Authors in earlier eras made frequent use of Latin abbreviations [Rydberg-Cox 2003]; these abbreviations, such as *ibid.* and *opera cit.,* have persisted in humanities scholarship as shorthand for previously cited works in earlier notes. Reference parsing methodologies utilized in bioinformatics for reference parsing are therefore inadequate for the diversity and plurality confronting the task of humanities reference parsing.

Given that bibliographic references contain sequences of elements, their parsing has often been approached as a sequence modeling task [Chen et al 2012]. Hetzner applied Hidden Markov Models (HMM) [Hetzner 2008]. Peng and McCullum pioneered CRFs for this task [Peng and McCallum 2006]. Colavizza, Kaplan, and Romanello built on this work, developing a custom historiographical dataset and examining the performance of CRF-based classifiers [Colavizza and Kaplan 2015; Colavizza and Romanello 2017, Colavizza, Romanello, Kaplan

2018]. Alves et al also utilized this dataset to create a BiLSTM- CRF neural network-based model [Alves et al 2018].

3.4 Discussion

Reference extraction and reference parsing in the humanities both have been studied by researchers in computer science. The studies noted above have examined the feasibility of applying supervised machine learning models to the extraction of the references and the disambiguation of metadata within elements. However, the properties of bibliographic references have not been considered as the features for supervised models with the goal of predicting the type of bibliographic footnote. This strikes us as a missed opportunity. The ability to predict the type of footnote prior to the reference parsing task would enable more nuanced application of models. Moreover, leveraging grammatical content of the references themselves and their context within author commentary in the footnote strings would facilitate improved analysis. The experimental phase of this thesis discussed in subsequent chapters will conduct a performance analysis of different supervised machine learning models using features traditionally used in reference extraction and grammatical features to ascertain the feasibility footnote classification in scholarly, archival humanities data.

This chapter has defined bibliographic references, differentiated scientific and humanistic modes of scholarly citation, and discussed existing methodologies to the essential tasks of extraction and parsing. In the next chapter, we will discuss the methodology and dataset utilized in this thesis to classify footnotes containing bibliographic references and scholarly commentary using supervised machine learning methods and novel, grammatical features.

Chapter 4: Methodology

4.1 Dataset Construction

The experimental phase of this thesis examines bibliographic references through a novel approach. Rather than concentrating on reference extraction or parsing of individual citations as prior work has done, we focus on references and scholarly commentary embedded in footnotes drawn from archival data. Our analysis treats the entire footnote as the unit of analysis. Our purpose is to examine the feasibility of analyzing the text of these footnotes in order to classify them according to the following categories given in Table 4-1.

Table 4-1

Bibliographic Footnote Category	Description
1	Footnotes containing solely commentary and no citations.
2	Footnotes consisting of a single bibliographic reference
3	Footnotes consisting of multiple references and scholarly commentary
4	Footnotes consisting of multiple bibliographic references with no scholarly commentary
5	Footnotes consisting of single references and scholarly commentary

Existing datasets of bibliographic references from the field of history were deemed unsuitable for this novel task. Therefore, we developed our own dataset of humanities scholarship derived from the JSTOR archive for this experiment [JSTOR].

JSTOR in the humanities disciplines is analogous to IEEE Explore and the ACM Digital Library for the fields of software engineering and computer science. It includes thousands of humanities journals and monographs published from the nineteenth century onward. Using issues from the *Journal of American History* (*JAH)* from the years 1964 and 1966, we extracted 495 footnotes from 8 research articles covering a diverse range of topics and periods in the history of

the United States and the North American continent to create the corpus. This process was completed manually to ensure accuracy of the data. Plaintext files corresponding to each research article were created. Footnotes were extracted from the article PDFs as text strings into these text files. These footnote strings were visually compared with the pages of the PDFs for accuracy and minor OCR errors detected by this manual inspection process were corrected. Table 4-1 shows the distribution of footnotes by article in the JSTOR *Journal of American History (JAH)* dataset.

Table 4-2

Article ID (in JSTOR database)	**Title**	**Number of Footnotes**
1887566	*The Historian as Editor: Francis Parkman's Reconstruction of Sources in Montcalm and Wolfe*	39
1887567	*Josiah Strong and American Nationalism: A Reevaluation*	70
1887568	*Progressives and the Great Society*	42
1887569	*Drafting the NRA Code of Fair Competition for the Bituminous Coal Industry*	95
1887570	*Hull, Russian Subversion in Cuba, and Recognition of the U.S.S.R*	48
1887571	*The Army "Mutiny" of 1946*	70
1888010	*Charles W. Eliot, University Reform, and Religious Faith in America, 1869-1909*	70
1888011	*The Massachusetts State Texas Committee: A Last Stand Against the Annexation of Texas*	61

4.2 Feature Extraction and Preprocessing

The plain text files created during the initial phase of dataset creation were then mined to extract the features for training the machine learning classifiers. Two sets of features were

extracted from the same set of footnote strings. The first set of features were derived from [Councill et al 2008]'s work on citation extraction from "reference strings." The 7 features selected are:

- String length in characters
- Title case word frequency
- Upper case word frequency
- Lower case word frequency
- Punctuation character
- Word count
- Number count

The second set of features are novel in this domain of bibliographic reference analysis. No prior work in this field has examined the suitability of using grammatical features for bibliographic reference classification in footnote strings. Therefore, we tokenized the footnote strings and applied part-of-speech tagging from the domain of natural language processing to these tokenized strings. For the parts-of-speech categories, we utilized the standard Penn Treebank tagset [Marcus et al 1993]. Minimal pre-processing was applied to the footnote strings for the second feature set to remove hyphenated number tokens. Tokenization and part-of-speech classification for the tokenized strings were performed using the Natural Language Toolkit [Bird 2006]. Tagging used the NLTK's built-in Perceptron tagger pre-trained on the Penn Treebank corpus. Figures 4-1 and 4-2 show the process of feature extraction for each set of features.

Figure 4-1

Process for extracting traditional features

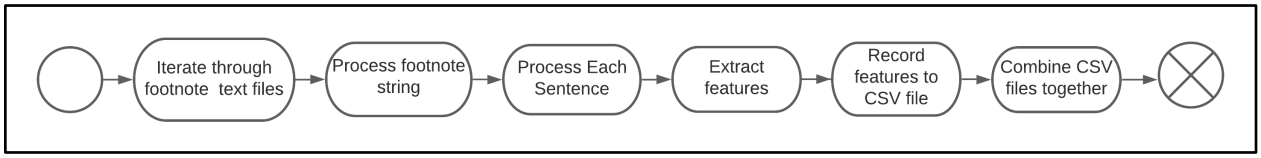
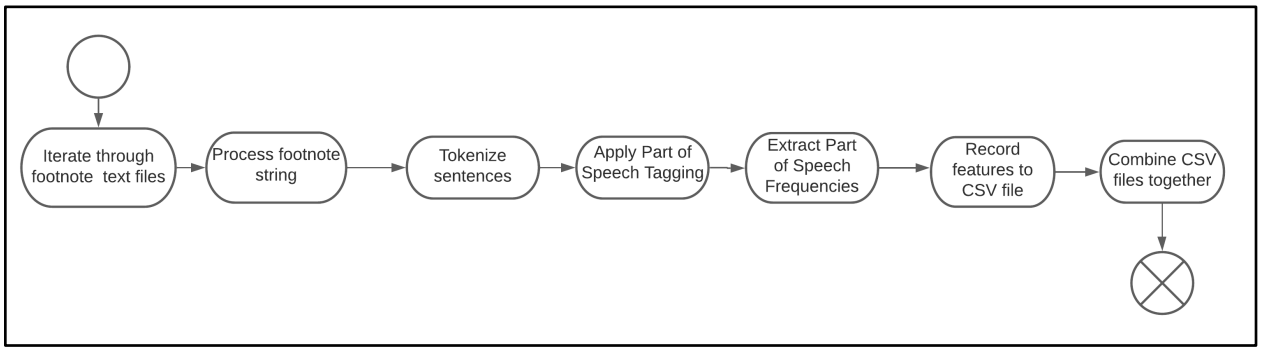


Figure 4-2

Process for applying part of speech tagging and extracting grammatical features



Both feature sets are exclusively numeric in the sense that they are frequencies of particular citation and semantic elements appearing in the footnote strings (as opposed to binary or categorical data) [Kelleher et al 2015]. For reference, the first set of extracted features from JSTOR *JAH* dataset is given in the Appendix.

4.3 Supervised Machine Learning Models

The experimental phase of this thesis was approached as a supervised machine learning problem, particularly a multiple-classification problem. We wanted to examine single and hybrid supervised machine learning methodologies and compare their performance on these two diverse feature sets. Therefore, we trained a suite of standard single and hybrid machine learning classifiers using the JSTOR *JAH* Dataset to predict *the bibliographic footnote category*, including:

- Multinomial Naïve Bayes

- K-Nearest Neighbors
- Decision Tree
- Gradient Boosting
- Stochastic Gradient Descent
- Logistic Regression
- Random Forest
- Support Vector Machine

The choice of models was in part determined by the nature of the features in the dataset, i.e. Multinomial Naïve Bayes was selected rather than Bernouli or Gaussian Naïve Bayes.

4.4 Tools and Software Libraries

In the spirit of software engineering that seeks to leverage existing software libraries and tools in conjunction to solve problems, we utilized the implementations of the machine learning algorithms provided by the Sci-Kit Learn Library [Pedregosa et al 2011]. Feature extraction for both feature sets was accomplished using a custom module of our own design. The figures demonstrating classifier performance and the confusion matrices were created using the Matplotlib and Seaborn libraries [Hunter 2007, Waskom 2021]. The code for extraction of both the traditional and grammatical features as well as integrating the classifiers was written in Python targeting Python version 3.7. The experimental scenarios discussed in the subsequent chapter were run on an Apple iMac with a 3.6 GHz Intel Core i9 processor with 16 GB of RAM.

Chapter 5: Experimental Results

This thesis's experiment is structured in two supervised machine learning scenarios. Scenario 1 trains a suite of single and hybrid classifiers on a set of traditional features originally utilized in bibliographic reference extraction. Scenario 2 trains the same set of classifiers using grammatical features drawn from the part-of-speech frequencies from bibliographic footnotes in historiographic archival data. This chapter discusses the results of both scenarios then concludes with experiment evaluation.

5.1 Scenario 1: Bibliographic Reference Classification using Traditional Features

Scenario 1 trains a suite of standard, supervised machine learning classifiers using features extracted from the JSTOR *JAH* Dataset discussed in the preceding chapter. This problem was structured as a multi-class, supervised learning problem. These classifiers were trained to predict the category of a given historiographic footnote based on the properties of its citations. The dataset was split into 75% training and 25% testing. Each classifier was trained using identical splits of the dataset into training and testing data.

Each classifier was trained and evaluated twice for the scenario, using unmodified features and with feature-scaling applied [Han et al 2011]. Standard Scaling was applied to the features prior to training Decision Tree, Gradient Boosting, K-Nearest Neighbors, Logistic Regression, Stochastic Gradient Descent, and Support Vector Machine classifiers. Min-Max Scaling was applied to the Multinomial Naïve Bayes feature samples. The accuracy of the classifiers for each round of the scenario are given in Figures 5-1 and 5-2.

Figure 5-1

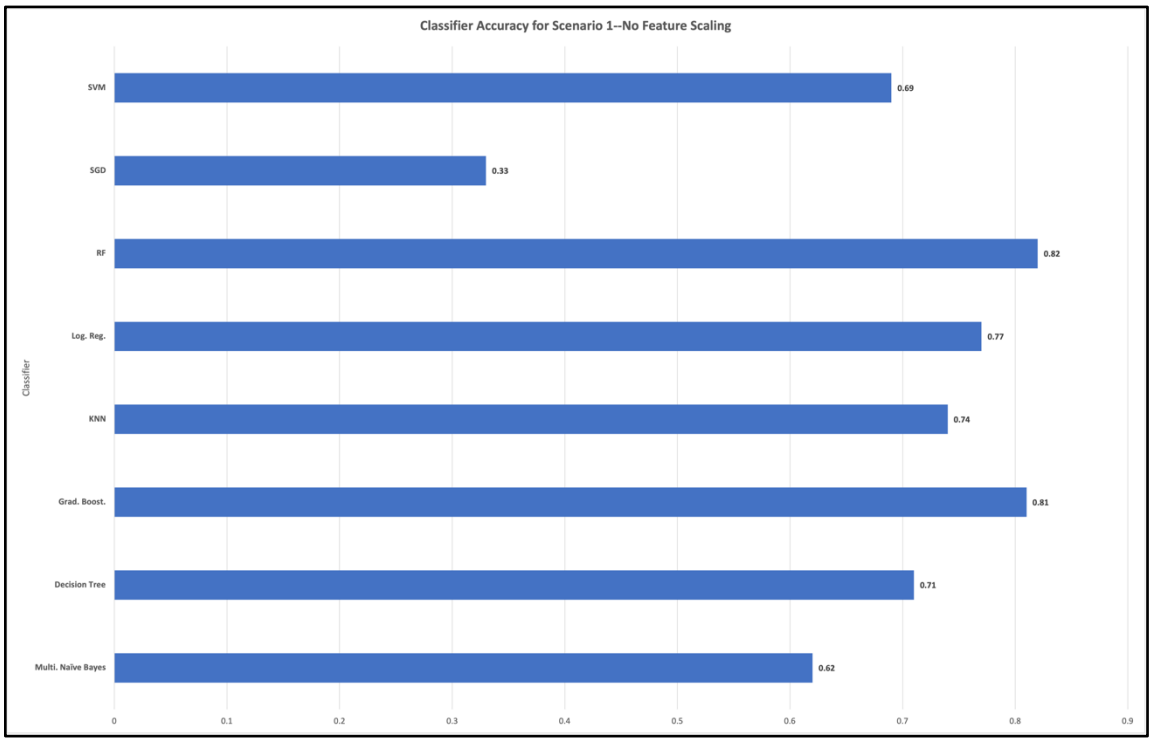
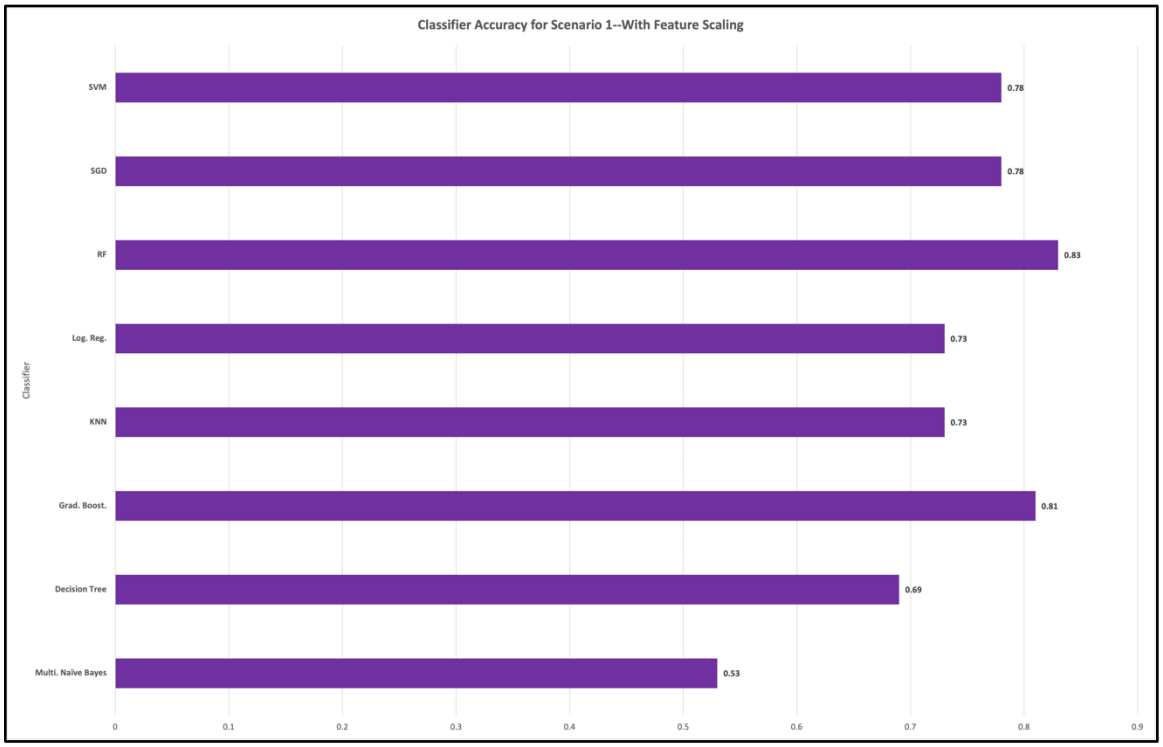


Figure 5-2



5.2 Scenario 2: Bibliographic Reference Classification using Grammatical Features

Scenario 2 trains the same set of supervised single and hybrid classifiers used in the first scenario with features extracted from the JSTOR *JAH* Dataset discussed in the preceding chapter. Using the features discussed in Section 4.2 above, the suite of classifiers were trained to predict the category of a given historiographic footnote based on the grammatical characteristics of its citations. As in Scenario 1, the dataset was split into 75% training and 25% testing. The same split of the dataset was used to train and evaluate each classifier.

Each classifier was trained and evaluated twice for the scenario, using unmodified features and with feature-scaling applied. The accuracies of the classifiers for each round of Scenario 2 are given in Figures 5-3 and 5-4.

Figure 5-3

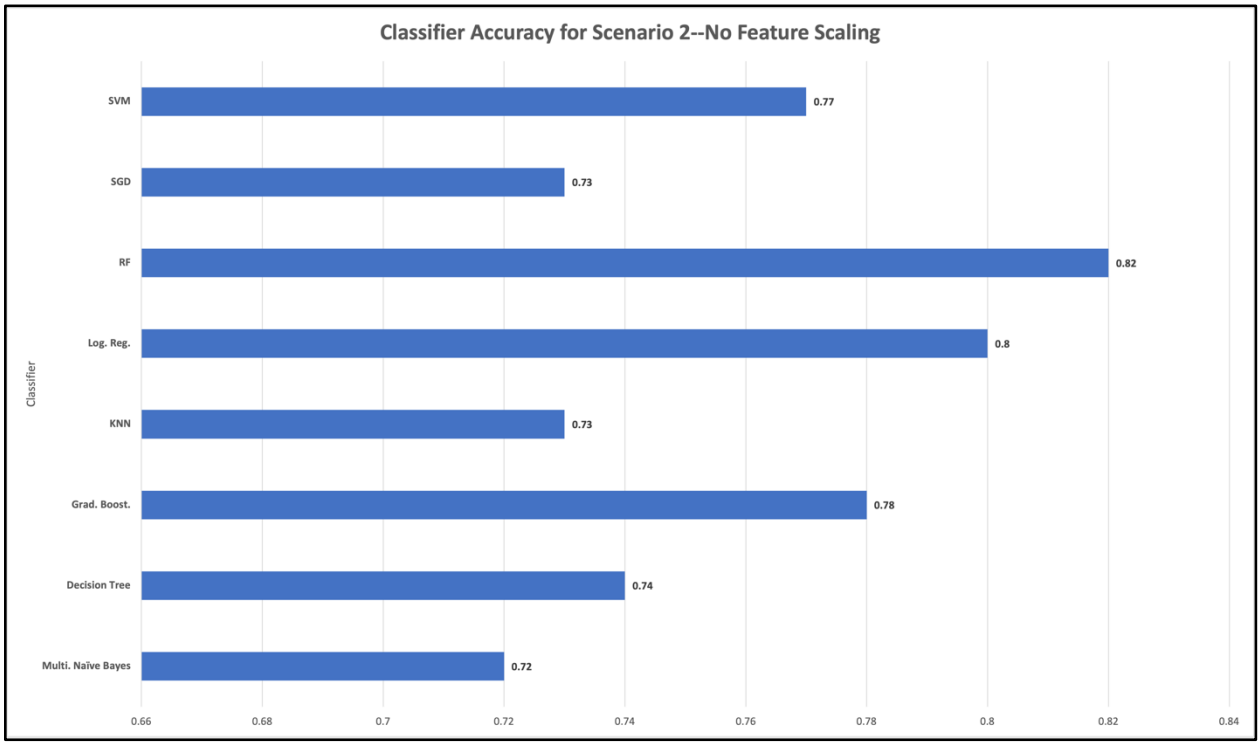
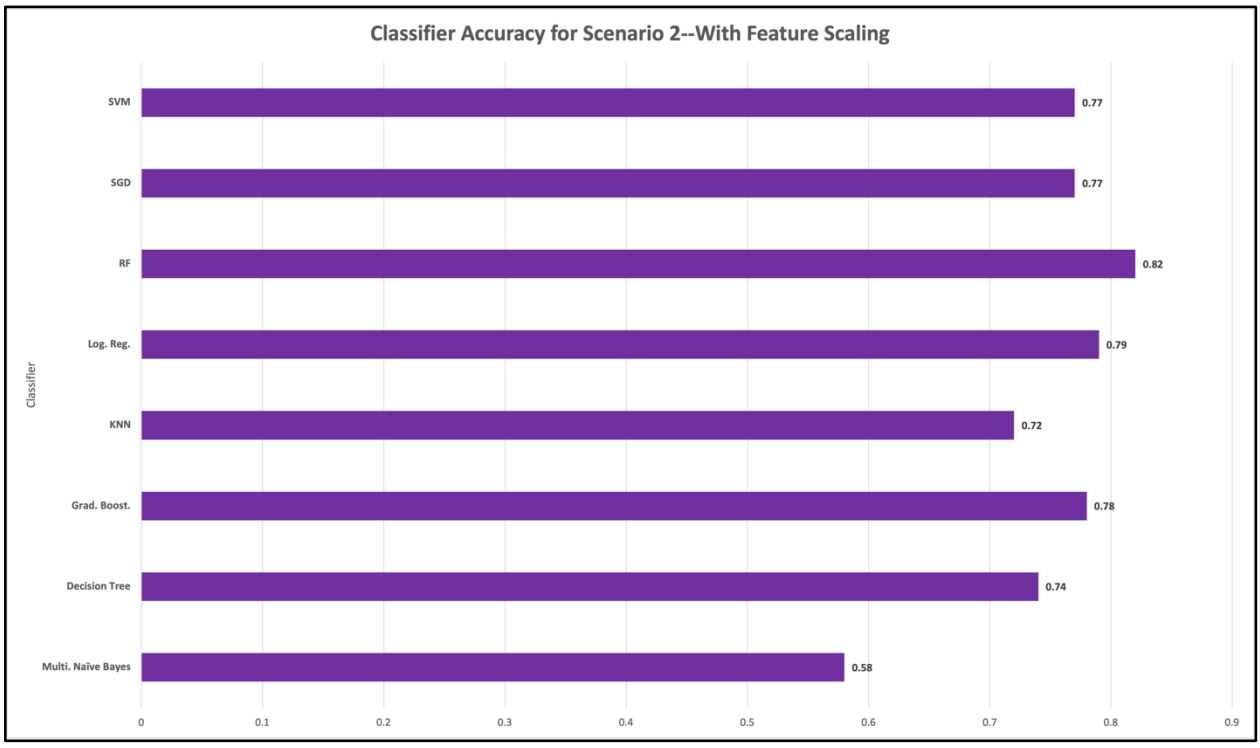


Figure 5-4



5.3 Discussion

For Scenario 1 using traditional features for training, the best performing single classifiers were Logistic Regression with an accuracy of 77% (no feature scaling) and a tie with an accuracy of 78% between Stochastic Gradient Descent and Support Vector Machine with feature scaling applied. The best performing hybrid classifier and best performing classifier overall was Random Forest with an accuracy of 83%.

For Scenario 2 using the novel grammatical features, the best performing single classifier was Logistic Regression with an accuracy of 80% without feature scaling. The worst performing classifier was Multinomial Naïve Bayes with an accuracy of 58% with feature scaling. As with Scenario 1, the best performing hybrid classifier was Random Forest with an accuracy of 82%.

The near tie in performance for the Random Forest classifier for both scenarios with an 82% vs 83% accuracy makes sense. Despite the different feature sets, traditional features and

grammatical features capture some similar properties of the data. For example, the frequencies of proper and common nouns in the data will be similar to the frequencies of upper and lowercase words. Where the grammatical features showed particular differentiation from the traditional ones in classifier performance was in the cases of the Multinomial Naïve Bayes, Decision Tree, Gradient Boosting, Stochastic Gradient Descent, and Support Vector Machine classifiers. Using grammatical features for training without feature-scaling improved the accuracy of Multinomial Naïve Bayes by 10%, Decision Tree by 3%, and Support Vector Machine by 6%.

Feature Scaling in both scenarios yielded mixed results. Without feature scaling, the logistic regression classifier's model would not fully converge. For Scenario 1, feature scaling led to a 45% improvement in accuracy for Stochastic Gradient Descent and a 9% improvement in accuracy for Support Vector Machine. In contrast, scaling yielded marginal improvement results in Scenario 2. The primary improvement was a 4% performance improvement for Stochastic Gradient Descent.

Several conclusions can be drawn from these results. First, hybrid machine learning models, particularly Random Forest, yields the best results regardless of if traditional or grammatical features are utilized for training. Furthermore, grammatical features show particular promise for improving the accuracy of classifiers such as Multinomial Naïve Bayes, Stochastic Gradient Descent, and Support Vector Machine. Finally, this experiment demonstrates the overall utility of grammatical features for machine learning problems of this type.

5.4 Evaluation

To evaluate the quality of the models built by the classifiers in these scenarios, ten-fold cross-validation was performed. Likewise, confusion matrices were generated for each model to

identify mis-classifications. Figures 5-5 and 5-6 show comparisons of individual classifier accuracy for both scenarios compared with the average cross-validation accuracy.

Figure 5-5

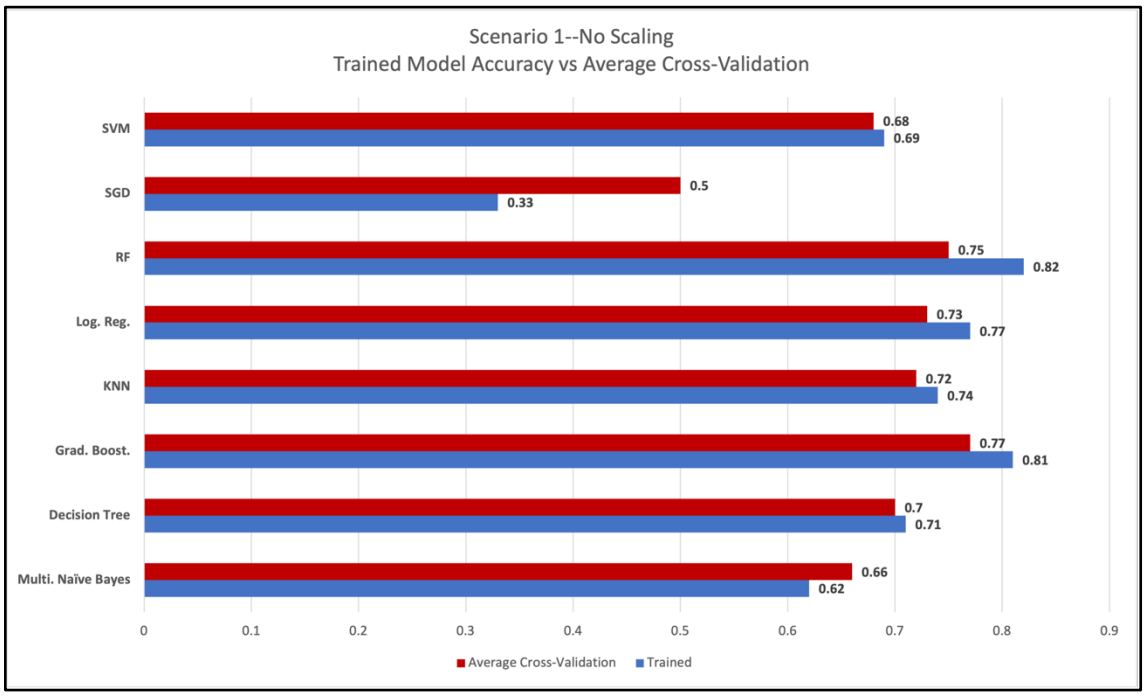


Figure 5-6

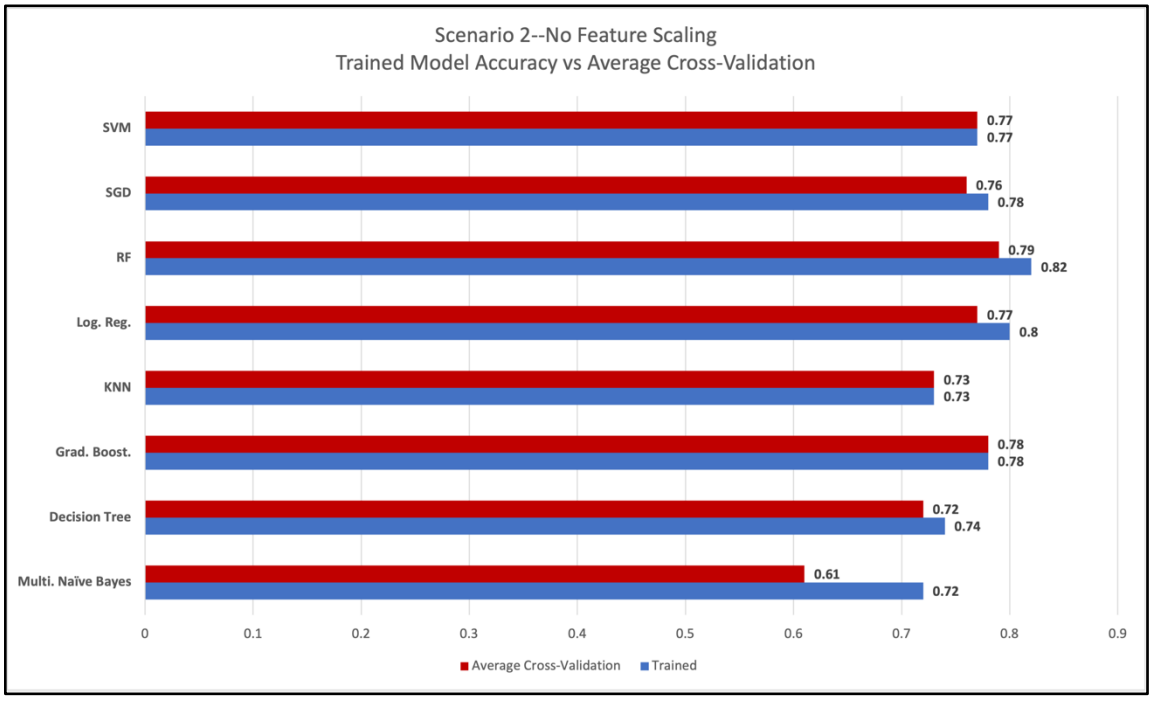


Figure 5-7

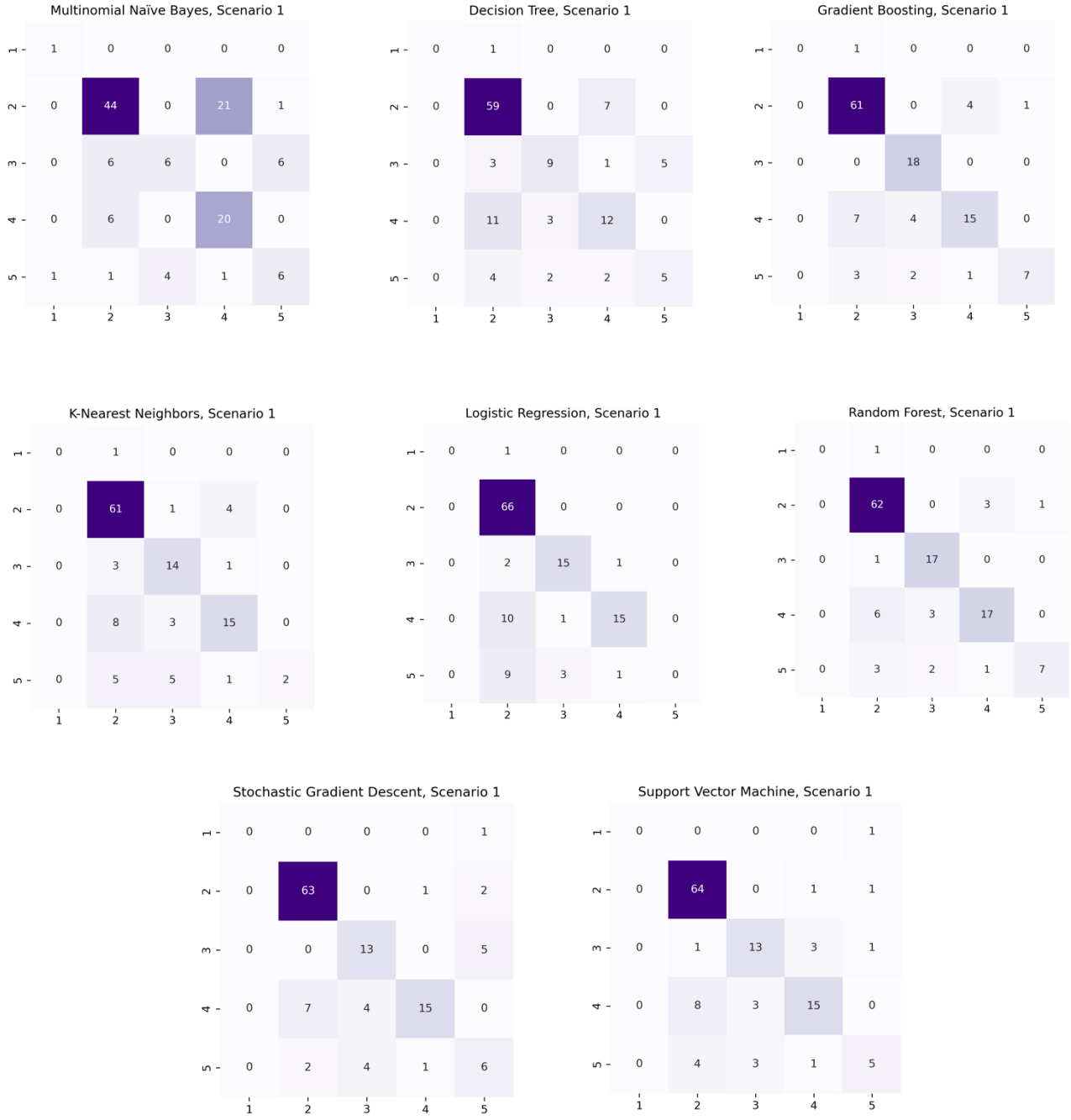
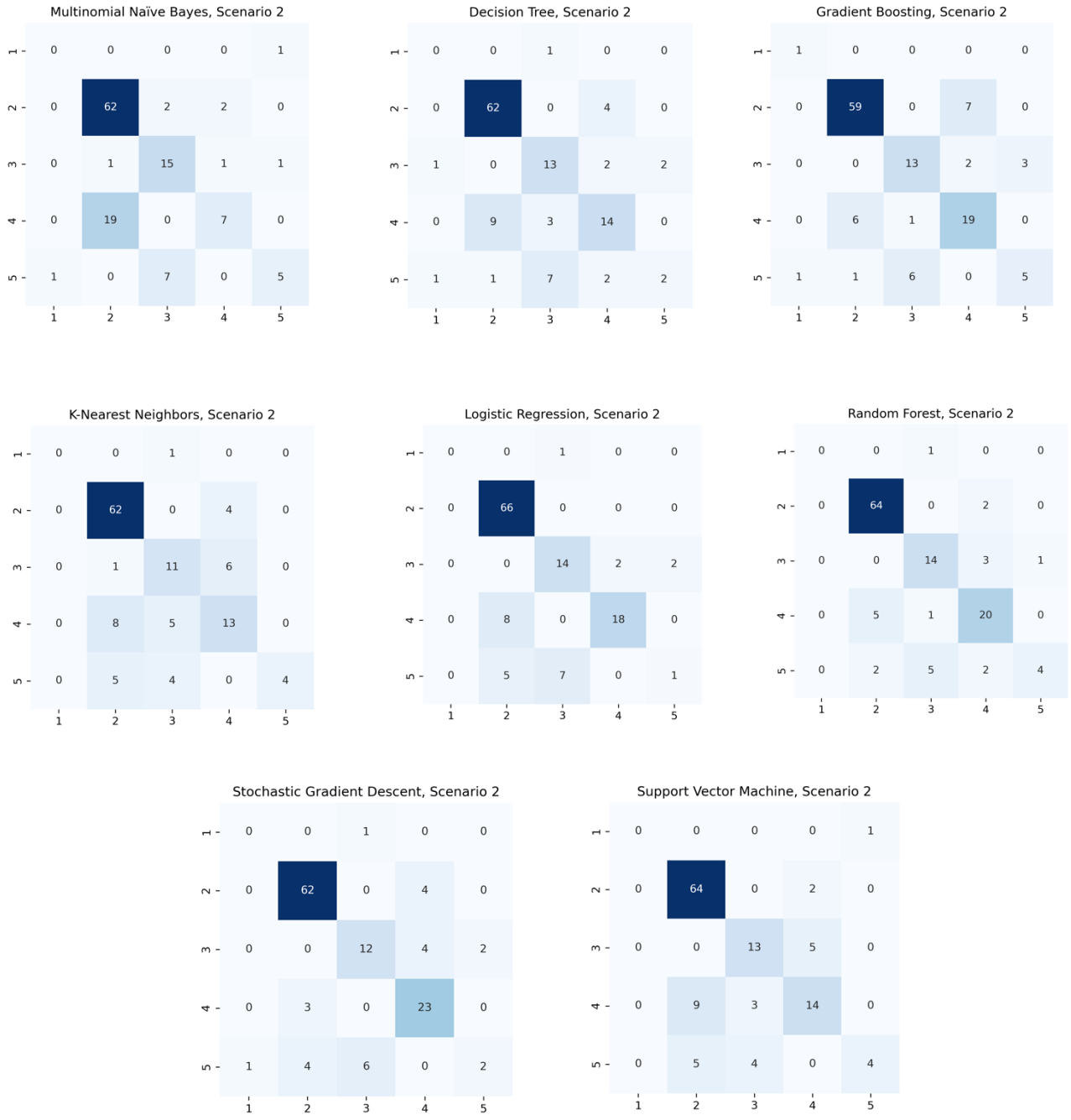

Multinomial Naïve Bayes, Scenario 1
1 1 0 0 0 0
2 0 44 0 21 1
3 0 6 6 0 6
4 0 6 0 20 0
5 1 1 4 1 6
1 2 3 4 5
Decision Tree, Scenario 1
1 0 1 0 0 0
2 0 59 0 7 0
3 0 3 9 1 5
4 0 11 3 12 0
5 0 4 2 2 5
1 2 3 4 5
Gradient Boosting, Scenario 1
1 0 1 0 0 0
2 0 61 0 4 1
3 0 0 18 0 0
4 0 7 4 15 0
5 0 3 2 1 7
1 2 3 4 5
K-Nearest Neighbors, Scenario 1
1 0 1 0 0 0
2 0 61 1 4 0
3 0 3 14 1 0
4 0 8 3 15 0
5 0 5 5 1 2
1 2 3 4 5
Logistic Regression, Scenario 1
1 0 1 0 0 0
2 0 66 0 0 0
3 0 2 15 1 0
4 0 10 1 15 0
5 0 9 3 1 0
1 2 3 4 5
Random Forest, Scenario 1
1 0 1 0 0 0
2 0 62 0 3 1
3 0 1 17 0 0
4 0 6 3 17 0
5 0 3 2 1 7
1 2 3 4 5
Stochastic Gradient Descent, Scenario 1
1 0 0 0 0 1
2 0 63 0 1 2
3 0 0 13 0 5
4 0 7 4 15 0
5 0 2 4 1 6
1 2 3 4 5
Support Vector Machine, Scenario 1
1 0 0 0 0 1
2 0 64 0 1 1
3 0 1 13 3 1
4 0 8 3 15 0
5 0 4 3 1 5
1 2 3 4 5

Figure 5-8


Multinomial Naïve Bayes, Scenario 2
0 0 0 0 1
0 62 2 2 0
0 1 15 1 1
0 19 0 7 0
1 0 7 0 5
Decision Tree, Scenario 2
0 0 1 0 0
0 62 0 4 0
1 0 13 2 2
0 9 3 14 0
1 1 7 2 2
Gradient Boosting, Scenario 2
1 0 0 0 0
0 59 0 7 0
0 0 13 2 3
0 6 1 19 0
1 1 6 0 5
K-Nearest Neighbors, Scenario 2
0 0 1 0 0
0 62 0 4 0
0 1 11 6 0
0 8 5 13 0
0 5 4 0 4
Logistic Regression, Scenario 2
0 0 1 0 0
0 66 0 0 0
0 0 14 2 2
0 8 0 18 0
0 5 7 0 1
Random Forest, Scenario 2
0 0 1 0 0
0 64 0 2 0
0 0 14 3 1
0 5 1 20 0
0 2 5 2 4
Stochastic Gradient Descent, Scenario 2
0 0 1 0 0
0 62 0 4 0
0 0 12 4 2
0 3 0 23 0
1 4 6 0 2
Support Vector Machine, Scenario 2
0 0 0 0 1
0 64 0 2 0
0 0 13 5 0
0 9 3 14 0
0 5 4 0 4

Figures 5-7 and 5-8 show confusion matrices for the best performing example of each classifier in both scenarios. A consistent pattern emerges from an analysis of the matrices. This pattern indicates that the classifiers struggled to distinguish footnotes containing a single citation from those that had multiple citations (categories 2 and 4). Additionally, due to the few samples for categories 1 and 5 (no commentary and single citation with commentary), additional training data will facilitate improved classifier performance in future work.

These two experimental scenarios demonstrate the viability of utilizing grammatical features to distinguish between bibliographic footnote types and the superiority of hybrid, supervised machine learning models such as Random Forest over single classifiers for this classification task. Additionally, the notable improved performance of models such as Multinomial Naïve Bayes using grammatical features indicates their independent utility for training compared to traditional feature types drawn from reference extraction tasks. While models constructed with grammatical features did not outperform the best models trained with traditional features, the performance improvements seen in the Multinomial Naïve Bayes and Decision Tree classifiers suggest grammatical features are complementary in nature to traditional, extractive features. Moreover, further experiments with a larger, more diverse dataset will be necessary to improve classifier performance on distinguishing footnote types such as single citation without commentary and notes with no citations.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

Bibliographic references are integral to scholarly discourse in humanities disciplines. While prior work has focused on reference extraction and parsing, little research has investigated the classification of footnotes containing bibliographic citations and author commentary using supervised machine learning methodologies. For this thesis, in Chapter 2 we contextualized bibliographic reference analysis within the broader domain of archival document processing through an original literature survey of current techniques, tools, and trends in the field of historical document processing. In Chapter 3, we reviewed related work on bibliographic citation extraction and machine learning reference parsing techniques. Finally, using a historiographic dataset drawn from the JSTOR humanities archive as discussed in Chapter 4, in Chapter 5 we trained and compared the performance of a suite of single and hybrid machine learning classifiers on a novel, previously unexplored bibliographic reference classification task. Moreover, as a part of this analysis, we compare the performance of traditional features and novel, grammatical features drawn from natural language processing.

Our work in this thesis demonstrates the superiority of hybrid models for classification of scholarly footnotes containing historiographic bibliographic references, the transferability of features from reference extraction to this research problem, and the viability of training machine learning models for this task utilizing novel, grammatical features.

6.2 Future Work

This thesis elucidates that prior studies have given much focus to bibliographic reference extraction and the parsing of reference metadata. The classification of scholarly footnotes containing bibliographic citations examined in this work was envisioned as a middle step

following reference extraction and prior to citation parsing. Future work will be necessary to determine if knowledge of the type of scholarly footnote prior to reference parsing enhances the accuracy of parsing.

Moreover, since the classification task undertaken in this thesis was novel, a new dataset had to be created to accomplish it. Manually annotating the dataset for this research took significant time and the necessity of proceeding with the experimental phase of this thesis constrained its size. Therefore, future work could expand the size and scope of this dataset beyond the *Journal of American History* to include additional articles and additional historiographical genres. A study of differences in citation patterns between scholarly monographs and journal articles and a comparison of machine learning model performance would be of interest. With a sufficiently large dataset of reference strings, the performance of Big Data analytics architectures, such as Hadoop and Spark, could be studied.

Another novel aspect of this thesis was the use of grammatical features, particularly part-of-speech tags applied to the elements of in scholarly footnotes. Developing rich grammatical models of scholarly footnotes, including named-entity analysis and morphology could improve the accuracy and utility of grammatical features for reference string classification.

This thesis focused on comparing performance of single and hybrid supervised machine learning models. Additional hybrid models could be explored to see if they lead to improved model performance. For example, voting and other kinds of boosting techniques could be examined.

As demonstrated, our models achieved an average accuracy of 71.06% across both types of feature scaling using traditional reference extraction features and an 75.38% accuracy across both types of feature scaling using new grammatical features. Yet, the cross-validation

performance indicated that higher accuracies are obtainable with other partitions of the training and testing data. More detailed feature engineering and principal component analysis could yield insight into which features are most useful for bibliographic reference analysis.

This thesis has focused on supervised machine learning methodologies for bibliographic reference analysis. Yet, the opportunities of using computational methods and developing software systems to advance "archival analytics" are many. Perhaps French historian Emmanuel Le Roy Ladurie was correct when he wrote: "The historian of tomorrow will be a programmer …" [Ladurie 1979].

BIBLIOGRAPHY

Anderson, David P. 2009. "Biographies: Tom Kilburn: A Pioneer of Computer Design." *IEEE Annals of the History of Computing* 31, no. 2: 82–86. https://doi.org/10.1109/MAHC.2009.32.

Baechler, Micheal, and Rolf Ingold. 2010. "Medieval Manuscript Layout Model." In *Proceedings of the 10th ACM Symposium on Document Engineering*, 275–78. DocEng '10. Manchester, United Kingdom: Association for Computing Machinery. https://doi.org/10.1145/1860559.1860622.

Bamman, David, and David Smith. 2012. "Extracting Two Thousand Years of Latin from a Million Book Library." *Journal on Computing and Cultural Heritage* 5, no. 1: 2:1-2:13. https://doi.org/10.1145/2160165.2160167.

Ben Messaoud, Ines, Hamid Amiri, Haikal El Abed, and Volker Märgner. 2012. "Binarization Effects on Results of Text-Line Segmentation Methods Applied on Historical Documents." In *2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA)*, 1092–97. https://doi.org/10.1109/ISSPA.2012.6310453.

Bird, Steven. 2006. "NLTK: The Natural Language Toolkit." In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, 69–72. Sydney, Australia: Association for Computational Linguistics. https://doi.org/10.3115/1225403.1225421.

Bosch, Vicente, Alejandro Héctor Toselli, and Enrique Vidal. 2014. "Semiautomatic Text Baseline Detection in Large Historical Handwritten Documents." In *2014 14th International Conference on Frontiers in Handwriting Recognition*, 690–95. https://doi.org/10.1109/ICFHR.2014.121.

Bowles, Edmund A. 1967. *Computers in Humanistic Research: Readings and Perspectives*. Englewood Cliffs, N.J.

Breuel, Thomas M., Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. 2013. "High-Performance OCR for Printed English and Fraktur Using LSTM Networks." In *2013 12th International Conference on Document Analysis and Recognition*, 683–87. https://doi.org/10.1109/ICDAR.2013.140.

Bukhari, Syed Saqib, Ahmad Kadi, Mohammad Ayman Jouneh, Fahim Mahmood Mir, and Andreas Dengel. 2017. "AnyOCR: An Open-Source OCR System for Historical Archives." In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:305–10. https://doi.org/10.1109/ICDAR.2017.58.

Bukhari, Syed Saqib, Faisal Shafait, and Thomas M. Breuel. 2012. "An Image Based Performance Evaluation Method for Page Dewarping Algorithms Using SIFT Features." In *Camera-Based Document Analysis and Recognition*, edited by Masakazu Iwamura and

Faisal Shafait, 138–49. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-29364-1_11.

Busa, R. 1980. "The Annals of Humanities Computing: The Index Thomisticus." *Computers and the Humanities* 14, no. 2: 83–90.

Ceruzzi, Paul E. 2003. *A History of Modern Computing*. 2nd ed. History of Computing. Cambridge, MA, USA: MIT Press.

Chandna, Swati, Francesca Rindone, Carsten Dachsbacher, and Rainer Stotzka. 2016. "Quantitative Exploration of Large Medieval Manuscripts Data for the Codicological Research." In *2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV)*, 20–28. https://doi.org/10.1109/LDAV.2016.7874306.

Chen, Chien-Chih, Kai-Hsiang Yang, Chuen-Liang Chen, and Jan-Ming Ho. 2012. "BibPro: A Citation Parser Based on Sequence Alignment." *IEEE Transactions on Knowledge and Data Engineering* 24, no. 2: 236–50. https://doi.org/10.1109/TKDE.2010.231.

Christy, Matthew, Anshul Gupta, Elizabeth Grumbach, Laura Mandell, Richard Furuta, and Ricardo Gutierrez-Osuna. 2017. "Mass Digitization of Early Modern Texts With Optical Character Recognition." *Journal on Computing and Cultural Heritage* 11, no. 1: 6:1-6:25. https://doi.org/10.1145/3075645.

Clausner, C., S. Pletschacher, and A. Antonacopoulos. 2011. "Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments." In *2011 International Conference on Document Analysis and Recognition*, 48–52. https://doi.org/10.1109/ICDAR.2011.19.

Colavizza, Giovanni, and Frédéric Kaplan. 2016. "On Mining Citations to Primary and Secondary Sources in Historiography." In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-It 2015 : 3-4 December 2015, Trento*, edited by Cristina Bosco, Sara Tonelli, and Fabio Massimo Zanzotto, 94–99. Collana Dell'Associazione Italiana Di Linguistica Computazionale. Torino: Accademia University Press. http://books.openedition.org/aaccademia/1439.

Colavizza, Giovanni, and Matteo Romanello. 2017. "Annotated References in the Historiography on Venice: 19th–21st Centuries." *Journal of Open Humanities Data* 3, no. November: 2. https://doi.org/10.5334/johd.9.

Colavizza, Giovanni, Matteo Romanello, and Frédéric Kaplan. 2018. "The References of References: A Method to Enrich Humanities Library Catalogs with Citation Data." *International Journal on Digital Libraries* 19, no. 2–3: 151–61. https://doi.org/10.1007/s00799-017-0210-1.

Councill, Isaac, C. Lee Giles, and Min-Yen Kan. 2008. "ParsCit: An Open-Source CRF Reference String Parsing Package." In *Proceedings of the Sixth International Conference on*

Language Resources and Evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA).

Edelstein, Dan, Paula Findlen, Giovanna Ceserani, Caroline Winterer, and Nicole Coleman. 2017. "Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project." *The American Historical Review*122, no. 2: 400–424. https://doi.org/10.1093/ahr/122.2.400.

Fischer, Andreas, Micheal Baechler, Angelika Garz, Marcus Liwicki, and Rolf Ingold. 2014. "A Combined System for Text Line Extraction and Handwriting Recognition in Historical Documents." In *2014 11th IAPR International Workshop on Document Analysis Systems*, 71–75. https://doi.org/10.1109/DAS.2014.51.

Fischer, Andreas, Horst Bunke, Nada Naji, Jacques Savoy, Micheal Baechler, and Rolf Ingold. 2012. "The HisDoc Project. Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries." https://doi.org/10.13140/2.1.2180.3526.

Fischer, Andreas, Volkmar Frinken, Alicia Fornés, and Horst Bunke. 2011. "Transcription Alignment of Latin Manuscripts Using Hidden Markov Models." In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing - HIP '11*, 29. Beijing, China: ACM Press. https://doi.org/10.1145/2037342.2037348.

Fischer, Andreas, Emanuel Indermühle, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. 2010. "Ground Truth Creation for Handwriting Recognition in Historical Documents." In *Proceedings of the 8th IAPR International Workshop on Document Analysis Systems - DAS '10*, 3–10. Boston, Massachusetts: ACM Press. https://doi.org/10.1145/1815330.1815331.

Fischer, Andreas, Emanuel Indermuhle, Volkmar Frinken, and Horst Bunke. 2011. "HMM-Based Alignment of Inaccurate Transcriptions for Historical Documents." In *2011 International Conference on Document Analysis and Recognition*, 53–57. https://doi.org/10.1109/ICDAR.2011.20.

Fischer, Andreas, Andreas Keller, Volkmar Frinken, and Horst Bunke. 2012. "Lexicon-Free Handwritten Word Spotting Using Character HMMs." *Pattern Recognition Letters*, Special Issue on Awards from ICPR 2010, 33, no. 7: 934–42. https://doi.org/10.1016/j.patrec.2011.09.009.

Fischer, Andreas, Kaspar Riesen, and Horst Bunke. 2010. "Graph Similarity Features for HMM-Based Handwriting Recognition in Historical Documents." In *2010 12th International Conference on Frontiers in Handwriting Recognition*, 253–58. https://doi.org/10.1109/ICFHR.2010.47.

Fischer, Andreas, Markus Wuthrich, Marcus Liwicki, Volkmar Frinken, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. 2009. "Automatic Transcription of Handwritten Medieval

Documents." In *2009 15th International Conference on Virtual Systems and Multimedia*, 137–42. https://doi.org/10.1109/VSMM.2009.26.

Frinken, Volkmar, Andreas Fischer, Markus Baumgartner, and Horst Bunke. 2014. "Keyword Spotting for Self-Training of BLSTM NN Based Handwriting Recognition Systems." *Pattern Recognition*, Handwriting Recognition and other PR Applications, 47, no. 3: 1073–82. https://doi.org/10.1016/j.patcog.2013.06.030.

Frinken, Volkmar, Andreas Fischer, and Carlos-D. Martínez-Hinarejos. 2013. "Handwriting Recognition in Historical Documents Using Very Large Vocabularies." In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing - HIP '13*, 67. Washington, District of Columbia: ACM Press. https://doi.org/10.1145/2501115.2501116.

Gatos, Basilis, Georgios Louloudis, and Nikolaos Stamatopoulos. 2014. "Segmentation of Historical Handwritten Documents into Text Zones and Text Lines." In *2014 14th International Conference on Frontiers in Handwriting Recognition*, 464–69. https://doi.org/10.1109/ICFHR.2014.84.

Grafton, Anthony. 2011. *Worlds Made by Words: Scholarship and Community in the Modern West*. Cambridge, Mass.: Harvard Univ. Press.

Granell, Emilio, Edgard Chammas, Laurence Likforman-Sulem, Carlos-D. Martínez-Hinarejos, Chafic Mokbel, and Bogdan-Ionuţ Cîrstea. 2018. "Transcription of Spanish Historical Handwritten Documents with Deep Neural Networks." *Journal of Imaging* 4, no. 1: 15. https://doi.org/10.3390/jimaging4010015.

Gudivada, Venkat N., Ricardo Baeza-Yates, and Vijay V. Raghavan. 2015. "Big Data: Promises and Problems." *Computer* 48, no. 3: 20–23. https://doi.org/10.1109/MC.2015.62.

Gudivada, Venkat N., Dhana Rao, and Vijay V. Raghavan. 2016. "Renaissance in Database Management: Navigating the Landscape of Candidate Systems." *Computer* 49, no. 4: 31–42. https://doi.org/10.1109/MC.2016.115.

Gupta, Deepank, Bob Morris, Terry Catapano, and Guido Sautter. 2009. "A New Approach towards Bibliographic Reference Identification, Parsing and Inline Citation Matching." In *Contemporary Computing*, edited by Sanjay Ranka, Srinivas Aluru, Rajkumar Buyya, Yeh-Ching Chung, Sumeet Dua, Ananth Grama, Sandeep K. S. Gupta, Rajeev Kumar, and Vir V. Phoha, 93–102. Communications in Computer and Information Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-03547-0_10.

Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. *Data Mining: Concepts and Techniques*. Elsevier.

Heil, Jacob, and Todd Samuelson. 2013. "Book History in the Early Modern OCR Project, or, Bringing Balance to the Force." *Journal for Early Modern Cultural Studies* 13, no. 4: 90–103. https://doi.org/10.1353/jem.2013.0050.

Hetzner, Erik. 2008. “A Simple Method for Citation Metadata Extraction Using Hidden Markov Models.” In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL ’08*, 280. Pittsburgh PA, PA, USA: ACM Press. https://doi.org/10.1145/1378889.1378937.

Hunter, John D. 2007. “Matplotlib: A 2D Graphics Environment.” *Computing in Science Engineering* 9, no. 3: 90–95. https://doi.org/10.1109/MCSE.2007.55.

Jamshidi, Pooyan, Claus Pahl, Nabor C. Mendonça, James Lewis, and Stefan Tilkov. 2018. “Microservices: The Journey So Far and Challenges Ahead.” *IEEE Software* 35, no. 3: 24–35. https://doi.org/10.1109/MS.2018.2141039.

Jenckel, Martin, Syed Saqib Bukhari, and Andreas Dengel. 2016. “AnyOCR: A Sequence Learning Based OCR System for Unlabeled Historical Documents.” In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 4035–40. https://doi.org/10.1109/ICPR.2016.7900265.

Jones, Steven E. 2018. *Roberto Busa, S.J., and the Emergence of Humanities Computing: The Priest and the Punched Cards*. New York: Routledge.

“JSTOR.” 2021. Accessed October 27, 2021. https://www.jstor.org/.

Kahle, Philip, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. “Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents.” In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 04:19–24. https://doi.org/10.1109/ICDAR.2017.307.

Ladurie, Emmanuel Le Roy. 1979. The Territory of the Historian. University of Chicago Press.

Le Bourgeois, F., and H. Emptoz. 2007. “DEBORA: Digital Access to Books of the Renaissance.” *International Journal of Document Analysis and Recognition (IJDAR)* 9, no. 2: 193–221. https://doi.org/10.1007/s10032-006-0030-0.

Lewis, C.S. 1970. “On the Reading of Old Books.” In *God in the Dock*, 217–25. Grand Rapids: Wm. B. Eerdmans.

Lopez, Patrice. 2009. “GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications.” In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, 473–74. ECDL’09. Corfu, Greece: Springer-Verlag.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. “Building a Large Annotated Corpus of English: The Penn Treebank.” *Computational Linguistics* 19, no. 2: 313–30.

Mas, Joan, Jose A. Rodriguez, Dimosthenis Karatzas, Gemma Sanchez, and Josep Llados. 2008. "HistoSketch: A Semi-Automatic Annotation Tool for Archival Documents." In *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, 517–24. https://doi.org/10.1109/DAS.2008.70.

McBurney, M.K., and P.L. Novak. 2002. "What Is Bibliometrics and Why Should You Care?" In *Proceedings. IEEE International Professional Communication Conference*, 108–14. https://doi.org/10.1109/IPCC.2002.1049094.

Ollagnier, Anaïs, Sébastien Fournier, and Patrice Bellot. 2016. "A Supervised Approach for Detecting Allusive Bibliographical References in Scholarly Publications." In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, 1–4. WIMS '16. Nîmes, France: Association for Computing Machinery. https://doi.org/10.1145/2912845.2912883.

———. 2018. "BIBLME RecSys: Harnessing Bibliometric Measures for a Scholarly Paper Recommender System." In *BIR 2018 Workshop on Bibliometric-Enhanced Information Retrieval*. Grenoble, France. https://hal.archives-ouvertes.fr/hal-01770588.

Papadopoulos, Christos, Stefan Pletschacher, Christian Clausner, and Apostolos Antonacopoulos. 2013. "The IMPACT Dataset of Historical Document Images." In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, 123–30. HIP '13. Washington, District of Columbia, USA: Association for Computing Machinery. https://doi.org/10.1145/2501115.2501130.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12, no. 85: 2825–30.

Peng, Fuchun, and Andrew McCallum. 2006. "Information Extraction from Research Papers Using Conditional Random Fields." *Information Processing & Management* 42, no. 4: 963–79. https://doi.org/10.1016/j.ipm.2005.09.002.

Philips, James, and Nasseh Tabrizi. 2020. "Historical Document Processing: A Survey of Techniques, Tools, and Trends:" In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 341–49. Budapest, Hungary: SCITEPRESS. https://doi.org/10.5220/0010177403410349.

Pintus, Ruggero, Ying Yang, and Holly Rushmeier. 2015. "ATHENA: Automatic Text Height Extraction for the Analysis of Text Lines in Old Handwritten Manuscripts." *Journal on Computing and Cultural Heritage* 8, no. 1: 1:1-1:25. https://doi.org/10.1145/2659020.

Pletschacher, Stefan, and Apostolos Antonacopoulos. 2010. "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework." In *2010 20th International Conference on Pattern Recognition*, 257–60. https://doi.org/10.1109/ICPR.2010.72.

Raha, Poulami, and Bhabatosh Chanda. 2019. “Restoration of Historical Document Images Using Convolutional Neural Networks.” In *2019 IEEE Region 10 Symposium (TENSYMP)*, 56–61. https://doi.org/10.1109/TENSYMP46218.2019.8971112.

Rahnemoonfar, Maryam, and Beth Plale. 2013. “Automatic Performance Evaluation of Dewarping Methods in Large Scale Digitization of Historical Documents.” In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 331–34. JCDL ’13. Indianapolis, Indiana, USA: Association for Computing Machinery. https://doi.org/10.1145/2467696.2467744.

Rath, Tony M., and R. Manmatha. 2007. “Word Spotting for Historical Documents.” *International Journal of Document Analysis and Recognition (IJDAR)* 9, no. 2: 139–52. https://doi.org/10.1007/s10032-006-0027-8.

Rodrigues Alves, Danny, Giovanni Colavizza, and Frédéric Kaplan. 2018. “Deep Reference Mining From Scholarly Literature in the Arts and Humanities.” *Frontiers in Research Metrics and Analytics* 3, no. July: 21. https://doi.org/10.3389/frma.2018.00021.

Roe, Edward, and Carlos A.B. Mello. 2013. “Binarization of Color Historical Document Images Using Local Image Equalization and XDoG.” In *2013 12th International Conference on Document Analysis and Recognition*, 205–9. https://doi.org/10.1109/ICDAR.2013.48.

Romanello, Matteo, Federico Boschetti, and Gregory Crane. 2009. “Citations in the Digital Library of Classics: Extracting Canonical References by Using Conditional Random Fields.” In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, 80–87. Suntec City, Singapore: Association for Computational Linguistics. https://aclanthology.org/W09-3610.

Romanello, Matteo, and Michele Pasin. 2013. “Citations and Annotations in Classics: Old Problems and New Perspectives.” In *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities*, 1–8. DH-CASE ’13. Florence, Italy: Association for Computing Machinery. https://doi.org/10.1145/2517978.2517981.

Rydberg-Cox, Jeffrey A. 2003. “Automatic Disambiguation of Latin Abbreviations in Early Modern Texts for Humanities Digital Libraries.” In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, 372–73. JCDL ’03. Houston, Texas: IEEE Computer Society.

———. 2009. “Digitizing Latin Incunabula: Challenges, Methods, and Possibilities.” *Digital Humanities Quarterly*003, no. 1.

Sastry, Panyam Narahari, and Ramakrishnan Krishnan. 2012. “A Data Acquisition and Analysis System for Palm Leaf Documents in Telugu.” In *Proceeding of the Workshop on Document Analysis and Recognition*, 139–46. DAR ’12. Mumbai, India: Association for Computing Machinery. https://doi.org/10.1145/2432553.2432578.

Scrivner, Buford. 1980. “Carolingian Monastic Library Catalogs and Medieval Classification of Knowledge.” *The Journal of Library History (1974-1987)* 15, no. 4: 427–44.

Serrano, Nicolas, Francisco Castro, and Alfons Juan. 2010. “The RODRIGO Database.” In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/477_Paper.pdf.

Shafait, Faisal. 2009. “Document Image Analysis with OCRopus.” In *2009 IEEE 13th International Multitopic Conference*, 1–6. https://doi.org/10.1109/INMIC.2009.5383078.

Sidebottom, Harry. 2005. “Roman Imperialism: The Changed Outward Trajectory of the Roman Empire.” *Historia: Zeitschrift Für Alte Geschichte* 54, no. 3: 315–30.

Simistira, Foteini, Mathias Seuret, Nicole Eichenberger, Angelika Garz, Marcus Liwicki, and Rolf Ingold. 2016. “DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts.” In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 471–76. https://doi.org/10.1109/ICFHR.2016.0093.

Springmann, Uwe, and Anke Lüdeling. 2017. “OCR of Historical Printings with an Application to Building Diachronic Corpora: A Case Study Using the RIDGES Herbal Corpus.” *Digital Humanities Quarterly* 011, no. 2.

Springmann, Uwe, Dietmar Najock, Hermann Morgenroth, Helmut Schmid, Annette Gotscharek, and Florian Fink. 2014. “OCR of Historical Printings of Latin Texts: Problems, Prospects, Progress.” In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 71–75. DATeCH ’14. Madrid, Spain: Association for Computing Machinery. https://doi.org/10.1145/2595188.2595205.

Su, Bolan, Shijian Lu, and Chew Lim Tan. 2010. “Binarization of Historical Document Images Using the Local Maximum and Minimum.” In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, 159–66. DAS ’10. Boston, Massachusetts, USA: Association for Computing Machinery. https://doi.org/10.1145/1815330.1815351.

Tabrizi, M. H. N. 2008. “Digital Archiving and Data Mining of Historic Document.” In *2008 International Conference on Advanced Computer Theory and Engineering*, 19–23. https://doi.org/10.1109/ICACTE.2008.220.

Terras, Melissa, James Baker, James Hetherington, David Beavan, Martin Zaltz Austwick, Anne Welsh, Helen O’Neill, Will Finley, Oliver Duke-Williams, and Adam Farquhar. 2018. “Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High-Performance Computing, and Transforming Access to British Library Digital Collections.” *Digital Scholarship in the Humanities* 33, no. 2: 456–66. https://doi.org/10.1093/llc/fqx020.

Tkaczyk, Dominika, Andrew Collins, Paraic Sheridan, and Joeran Beel. 2018. "Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers." In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 99–108. JCDL '18. Fort Worth, Texas, USA: Association for Computing Machinery. https://doi.org/10.1145/3197026.3197048.

Ul-Hasan, Adnan, Syed Saqib Bukhari, and Andreas Dengel. 2016. "OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters." In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 174–79. https://doi.org/10.1109/DAS.2016.51.

Vobl, Thorsten, Annette Gotscharek, Uli Reffle, Christoph Ringlstetter, and Klaus U. Schulz. 2014. "PoCoTo - an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts." In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 57–61. DATeCH '14. Madrid, Spain: Association for Computing Machinery. https://doi.org/10.1145/2595188.2595197.

Waskom, Michael L. 2021. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software* 6, no. 60: 3021. https://doi.org/10.21105/joss.03021.

Wei, Hao, Kai Chen, Anguelos Nicolaou, Marcus Liwicki, and Rolf Ingold. 2014. "Investigation of Feature Selection for Historical Document Layout Analysis." In *2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6. https://doi.org/10.1109/IPTA.2014.7001961.

Würsch, Marcel, Rolf Ingold, and Marcus Liwicki. 2017. "DivaServices—A RESTful Web Service for Document Image Analysis Methods." *Digital Scholarship in the Humanities* 32, no. suppl_1: i150–56. https://doi.org/10.1093/llc/fqw051.

Xie, Boya, Qin Ding, Hongjin Han, and Di Wu. 2013. "MiRCancer: A MicroRNA–Cancer Association Database Constructed by Text Mining on Literature." *Bioinformatics* 29, no. 5: 638–44. https://doi.org/10.1093/bioinformatics/btt014.

Yang, Ying, Ruggero Pintus, Enrico Gobbetti, and Holly Rushmeier. 2017. "Automatic Single Page-Based Algorithms for Medieval Manuscript Analysis." *Journal on Computing and Cultural Heritage* 10, no. 2: 9:1-9:22. https://doi.org/10.1145/2996469.

APPENDIX: JSTOR *JOURNAL OF AMERICAN HISTORY (JAH)* DATASET

Article ID	Footnote ID	Char. Count	Title Case Count	Upper Case Count	Lower Case Count	Punc. Count	Word Count	Num. Count	Bibl. Foot-note Category	Cit. Count
1887566	1	134	14	2	3	8	18	3	1	1
1887566	2	513	55	8	7	40	67	14	2	6
1887566	3	195	13	2	6	16	20	13	3	2
1887566	4	81	8	0	1	8	9	2	1	1
1887566	5	39	3	1	1	4	5	1	1	1
1887566	6	65	7	0	1	4	8	2	1	1
1887566	7	38	4	1	1	4	5	1	1	1
1887566	8	25	2	0	0	3	2	1	1	1
1887566	9	38	4	1	1	4	5	1	1	1
1887566	10	52	6	0	1	3	7	1	1	1
1887566	11	230	15	1	13	12	29	4	2	2
1887566	12	40	3	0	2	3	5	1	1	1
1887566	13	38	4	1	1	4	5	1	1	1
1887566	14	244	24	4	7	16	33	4	3	2
1887566	15	94	8	1	2	5	11	2	1	1
1887566	16	38	4	1	1	4	5	1	1	1
1887566	17	98	7	1	2	8	10	5	1	1
1887566	18	38	4	1	1	4	5	1	1	1
1887566	19	34	2	0	0	4	3	2	1	1
1887566	20	199	16	2	9	14	26	3	3	2
1887566	21	107	9	1	2	7	12	2	1	1
1887566	22	40	4	1	1	4	5	2	1	1
1887566	23	16	2	1	0	3	2	1	1	1
1887566	24	80	8	0	1	4	9	3	1	1
1887566	25	38	4	1	1	4	5	1	1	1
1887566	26	25	2	0	0	3	2	1	1	1
1887566	27	38	4	1	1	4	5	1	1	1
1887566	28	29	2	0	0	3	2	2	1	1
1887566	29	12	1	0	0	2	1	1	1	1

1887566	30	41	3	0	1	4	4	3	1	1
1887566	31	163	12	1	5	13	18	5	1	1
1887566	32	38	3	0	1	4	4	2	1	1
1887566	33	101	9	1	5	8	14	2	1	1
1887566	34	173	14	1	13	9	28	2	4	1
1887566	35	38	4	1	1	4	5	1	1	1
1887566	36	16	2	1	0	3	2	1	1	1
1887566	37	190	15	3	10	16	27	6	1	1
1887566	38	38	4	0	1	3	5	1	1	1
1887566	39	72	8	0	2	4	10	1	1	1
1887567	1	319	27	4	11	25	41	10	3	3
1887567	2	203	23	1	2	12	25	5	3	2
1887567	3	258	26	2	7	15	33	5	3	2
1887567	4	1371	112	6	61	84	177	42	2	11
1887567	5	172	22	0	2	11	24	6	3	2
1887567	6	32	4	0	0	3	4	2	1	1
1887567	7	22	3	0	0	3	3	1	1	1
1887567	8	323	17	1	31	15	49	6	2	2
1887567	9	564	28	8	37	41	73	22	2	10
1887567	10	37	2	1	0	5	3	3	1	1
1887567	11	98	7	2	1	8	10	4	3	2
1887567	12	507	35	9	19	43	62	20	2	9
1887567	13	148	13	0	0	24	13	17	3	7
1887567	14	794	47	4	21	43	71	48	2	2
1887567	15	44	3	1	1	4	5	2	1	1
1887567	16	665	55	3	27	39	85	17	2	6
1887567	17	354	26	2	20	24	47	8	2	4
1887567	18	35	3	0	0	4	3	3	1	1
1887567	19	418	21	1	25	23	47	9	4	1
1887567	20	68	6	0	1	8	7	4	3	2
1887567	21	48	5	0	0	6	5	4	3	2
1887567	22	93	10	0	0	8	10	4	3	2
1887567	23	324	22	1	12	19	37	8	4	1
1887567	24	93	11	0	0	10	11	3	3	2
1887567	25	467	34	4	23	28	58	12	2	5
1887567	26	279	20	2	25	14	45	5	2	2
1887567	27	38	2	0	0	5	2	4	1	1
1887567	28	267	26	3	5	26	33	10	3	4
1887567	29	33	2	0	0	4	2	3	1	1

1887567	30	266	22	0	14	21	36	7	2	3
1887567	31	27	3	0	0	3	3	1	1	1
1887567	32	285	28	1	4	22	33	13	3	4
1887567	33	40	5	0	0	5	5	2	3	2
1887567	34	28	3	0	0	4	3	2	1	1
1887567	35	16	1	0	0	2	1	2	1	1
1887567	36	13	1	0	0	2	1	1	1	1
1887567	37	128	11	1	5	11	17	4	3	2
1887567	38	399	20	0	37	12	57	6	4	1
1887567	39	71	8	0	0	6	8	3	1	1
1887567	40	95	4	0	11	5	15	0	4	1
1887567	41	344	12	0	33	14	47	5	4	1
1887567	42	36	2	0	0	4	2	4	1	1
1887567	43	86	7	0	2	8	9	7	3	2
1887567	44	25	2	0	0	3	2	1	1	1
1887567	45	156	10	0	10	10	20	5	2	4
1887567	46	135	12	1	3	11	16	4	3	2
1887567	47	289	26	0	10	23	36	9	2	4
1887567	48	1339	102	14	93	68	197	13	2	3
1887567	49	127	14	2	2	8	16	1	1	1
1887567	50	111	8	2	2	14	11	8	3	3
1887567	51	42	4	0	0	6	4	3	3	2
1887567	52	59	5	0	3	4	8	1	1	1
1887567	53	483	21	1	43	24	64	17	2	3
1887567	54	357	18	2	23	18	45	6	2	2
1887567	55	361	19	5	16	30	40	21	2	7
1887567	56	56	5	0	0	7	5	5	1	1
1887567	57	85	7	1	0	10	8	5	3	2
1887567	58	49	4	0	0	6	4	3	3	2
1887567	59	59	5	0	3	4	8	1	1	1
1887567	60	30	2	0	0	3	2	1	1	1
1887567	61	112	10	1	4	12	15	4	3	2
1887567	62	414	37	2	13	23	51	14	3	3
1887567	63	49	3	1	0	5	4	3	1	1
1887567	64	43	3	1	0	4	4	3	1	1
1887567	65	91	8	2	0	10	9	5	3	2
1887567	66	118	14	2	4	12	18	4	3	2
1887567	67	258	16	1	24	8	40	0	4	1
1887567	68	164	11	0	7	11	18	4	1	1

1887567	69	47	3	1	0	6	4	3	1	1
1887567	70	67	7	0	3	7	10	1	3	2
1887568	1	124	12	1	2	9	15	3	1	1
1887568	2	824	78	9	16	52	96	22	3	8
1887568	3	285	23	2	6	18	31	6	3	2
1887568	4	314	37	5	7	20	45	5	3	3
1887568	5	123	13	0	2	9	15	2	1	1
1887568	6	519	59	5	13	34	72	9	3	5
1887568	7	94	9	1	2	7	12	3	1	1
1887568	8	200	20	2	7	16	28	6	1	1
1887568	9	139	13	1	4	11	18	3	1	1
1887568	10	112	14	2	1	7	15	2	1	1
1887568	11	126	11	1	4	7	16	2	1	1
1887568	12	142	17	3	3	12	21	4	3	2
1887568	13	166	17	1	7	12	24	5	1	1
1887568	14	120	11	0	3	9	14	5	1	1
1887568	15	189	19	0	8	8	27	1	3	3
1887568	16	652	36	3	65	26	104	5	2	3
1887568	17	62	7	0	1	5	8	2	1	1
1887568	18	41	4	1	2	6	6	2	1	1
1887568	19	49	4	0	2	4	6	2	1	1
1887568	20	7	1	0	0	1	1	0	1	1
1887568	21	7	1	0	0	1	1	0	1	1
1887568	22	65	9	3	1	7	10	0	1	1
1887568	23	914	17	1	133	34	152	2	4	1
1887568	24	49	4	0	2	5	6	2	1	1
1887568	25	48	5	2	2	6	7	2	1	1
1887568	26	236	11	0	26	14	38	3	4	1
1887568	27	61	6	0	1	5	7	2	1	1
1887568	28	1067	44	7	114	50	164	10	2	3
1887568	29	59	4	1	2	7	7	3	1	1
1887568	30	44	3	1	0	5	4	4	1	1
1887568	31	51	4	0	1	5	6	2	1	1
1887568	32	198	7	0	24	6	31	1	4	1
1887568	33	38	4	1	0	5	4	3	1	1
1887568	34	51	5	0	1	5	6	2	1	1
1887568	35	41	3	0	2	4	5	2	1	1
1887568	36	924	50	7	99	39	150	5	2	3
1887568	37	553	18	2	71	21	90	4	2	2

1887568	38	168	6	0	17	7	23	2	4	1
1887568	39	45	4	0	2	4	6	2	1	1
1887568	40	38	3	0	2	5	5	2	1	1
1887568	41	107	13	1	2	6	15	1	1	1
1887568	42	51	5	0	1	4	6	2	1	1
1887569	1	801	50	5	68	46	121	17	2	5
1887569	2	281	22	2	9	17	33	6	3	3
1887569	3	580	31	3	53	26	88	6	2	2
1887569	4	133	14	2	1	9	16	6	3	2
1887569	5	45	2	1	1	5	4	4	1	1
1887569	6	71	7	1	0	7	8	5	3	2
1887569	7	40	3	1	0	4	4	3	1	1
1887569	8	230	20	3	5	26	25	10	3	5
1887569	9	154	16	5	3	14	21	4	1	1
1887569	10	32	3	1	0	3	4	1	1	1
1887569	11	100	6	0	9	8	15	3	4	1
1887569	12	214	20	2	6	11	26	4	1	1
1887569	13	188	16	3	9	9	27	3	1	1
1887569	14	37	3	1	0	4	4	2	1	1
1887569	15	50	5	0	1	6	6	3	1	1
1887569	16	87	10	0	0	5	10	1	1	1
1887569	17	131	16	2	1	10	17	3	1	1
1887569	18	150	15	2	4	13	21	2	1	1
1887569	19	50	4	0	2	4	6	1	1	1
1887569	20	37	4	0	0	4	4	1	1	1
1887569	21	275	7	0	34	9	41	1	4	1
1887569	22	475	17	0	58	18	75	6	4	1
1887569	23	51	5	0	1	5	6	3	1	1
1887569	24	83	12	3	1	8	13	2	1	1
1887569	25	137	16	3	4	10	21	2	1	1
1887569	26	73	8	0	1	5	9	2	1	1
1887569	27	96	10	2	2	7	13	2	1	1
1887569	28	74	7	0	2	5	9	3	1	1
1887569	29	54	5	0	1	5	6	4	1	1
1887569	30	74	8	2	2	7	10	3	1	1
1887569	31	360	35	5	16	22	52	9	3	2
1887569	32	50	5	0	1	6	6	3	1	1
1887569	33	66	6	3	1	6	9	3	1	1
1887569	34	53	5	0	1	5	6	4	1	1

1887569	35	211	23	3	4	13	27	8	3	2
1887569	36	50	5	0	1	3	6	1	1	1
1887569	37	126	14	2	5	12	19	3	3	2
1887569	38	92	8	0	3	8	11	4	1	1
1887569	39	36	4	0	0	5	4	2	1	1
1887569	40	29	3	0	0	3	3	1	1	1
1887569	41	82	9	1	2	6	11	4	1	1
1887569	42	69	10	3	1	8	11	2	1	1
1887569	43	38	4	0	0	6	4	2	1	1
1887569	44	78	7	1	3	8	10	4	1	1
1887569	45	34	4	0	0	5	4	1	1	1
1887569	46	106	13	0	3	6	16	3	1	1
1887569	47	57	6	0	0	6	6	4	3	2
1887569	48	152	15	0	2	10	17	5	3	2
1887569	49	89	9	0	1	8	10	4	3	2
1887569	50	88	10	0	0	12	10	5	3	3
1887569	51	27	2	0	0	3	2	2	1	1
1887569	52	61	6	3	1	6	9	2	1	1
1887569	53	205	11	2	20	11	33	5	4	1
1887569	54	42	2	1	1	6	4	4	1	1
1887569	55	31	4	0	0	4	4	2	1	1
1887569	56	7	1	0	0	1	1	0	1	1
1887569	57	57	7	0	0	7	7	3	3	2
1887569	58	90	9	1	0	13	10	6	3	3
1887569	59	67	7	1	0	8	8	4	3	2
1887569	60	147	11	2	2	17	15	9	3	3
1887569	61	494	24	2	44	31	71	14	2	5
1887569	62	158	11	1	6	17	18	9	3	3
1887569	63	42	3	0	2	4	5	1	1	1
1887569	64	393	11	2	51	15	62	1	4	1
1887569	65	50	5	1	0	4	6	0	1	1
1887569	66	32	2	1	0	4	3	2	1	1
1887569	67	60	6	1	0	4	7	0	1	1
1887569	68	23	2	0	1	3	3	0	1	1
1887569	69	55	5	0	3	6	8	2	1	1
1887569	70	16	1	0	1	2	2	0	1	1
1887569	71	182	21	6	5	20	27	4	3	3
1887569	72	206	9	0	20	11	29	3	4	1
1887569	73	159	15	0	4	12	19	3	3	2

1887569	74	509	15	0	57	16	72	5	2	2
1887569	75	117	9	2	1	16	12	9	3	3
1887569	76	237	13	2	17	15	33	6	2	2
1887569	77	396	21	1	34	24	56	12	2	4
1887569	78	69	7	1	0	8	8	4	3	2
1887569	79	139	12	1	2	16	15	8	3	4
1887569	80	188	16	3	4	19	23	9	3	3
1887569	81	182	16	1	5	14	22	6	3	3
1887569	82	103	9	1	3	6	13	5	3	2
1887569	83	233	16	5	3	22	24	13	3	4
1887569	84	691	28	1	88	33	117	4	2	2
1887569	85	69	6	1	3	2	10	2	1	1
1887569	86	156	13	2	1	20	17	10	3	5
1887569	87	168	12	3	5	18	21	8	3	4
1887569	88	77	8	0	0	12	8	6	3	2
1887569	89	44	5	0	0	7	5	4	1	1
1887569	90	202	17	3	2	22	22	10	3	4
1887569	91	59	2	2	3	4	7	2	1	1
1887569	92	75	5	0	4	8	9	3	3	2
1887569	93	64	6	1	3	6	9	2	1	1
1887569	94	121	12	1	0	11	13	4	3	2
1887569	95	149	15	1	4	9	19	3	2	2
1887570	1	211	22	0	5	17	27	9	3	3
1887570	2	264	25	3	3	23	29	11	3	3
1887570	3	105	11	2	0	8	12	2	1	1
1887570	4	113	11	0	3	8	14	4	1	1
1887570	5	147	13	0	5	6	18	4	1	1
1887570	6	156	13	1	7	10	21	2	4	1
1887570	7	102	9	4	5	11	15	4	1	1
1887570	8	444	16	1	51	19	68	5	2	2
1887570	9	140	12	2	4	11	17	4	3	2
1887570	10	183	11	3	8	14	21	5	2	3
1887570	11	72	8	1	0	6	8	2	1	1
1887570	12	95	11	1	1	7	12	3	1	1
1887570	13	172	17	2	3	10	22	3	1	1
1887570	14	209	16	0	6	12	23	5	3	2
1887570	15	130	9	1	7	8	16	5	1	1
1887570	16	646	36	3	53	27	92	8	2	3
1887570	17	777	38	7	49	43	89	10	2	10

1887570	18	137	9	1	5	9	15	2	1	1
1887570	19	34	1	0	0	6	1	7	1	1
1887570	20	12	1	0	0	2	1	1	1	1
1887570	21	238	3	0	34	8	37	3	4	1
1887570	22	85	8	0	1	7	9	2	1	1
1887570	23	375	19	1	30	14	50	6	4	1
1887570	24	22	3	0	0	2	3	1	1	1
1887570	25	64	3	1	1	5	5	2	2	3
1887570	26	184	7	1	21	8	29	3	4	1
1887570	27	105	6	0	4	5	10	2	1	1
1887570	28	91	4	0	5	5	9	2	1	1
1887570	29	37	2	0	0	5	2	2	1	1
1887570	30	59	5	1	2	5	7	2	1	1
1887570	31	60	4	0	2	5	6	2	1	1
1887570	32	60	3	0	0	7	4	2	1	1
1887570	33	93	4	0	5	5	9	2	1	1
1887570	34	42	2	0	0	4	2	4	1	1
1887570	35	396	21	1	22	26	43	10	2	4
1887570	36	123	8	2	4	12	14	5	3	2
1887570	37	674	27	0	69	30	99	6	2	3
1887570	38	90	8	1	2	8	10	4	1	1
1887570	39	127	4	2	3	6	11	5	1	1
1887570	40	123	13	0	3	8	16	2	3	2
1887570	41	286	27	1	9	24	37	9	3	3
1887570	42	80	7	1	0	6	7	3	1	1
1887570	43	35	3	0	0	5	3	2	1	1
1887570	44	297	4	0	44	9	48	0	0	0
1887570	45	24	2	0	1	3	3	0	1	1
1887570	46	378	25	3	33	19	59	5	4	1
1887570	47	69	9	2	0	5	9	3	1	1
1887570	48	304	14	0	30	8	44	4	2	2
1887571	1	211	22	0	5	17	27	9	4	1
1887571	2	264	25	3	3	23	29	11	1	1
1887571	3	105	11	2	0	8	12	2	1	1
1887571	4	113	11	0	3	8	14	4	1	1
1887571	5	147	13	0	5	6	18	4	1	1
1887571	6	156	13	1	7	10	21	2	4	1
1887571	7	102	9	4	5	11	15	4	1	1
1887571	8	444	16	1	51	19	68	5	1	1

1887571	9	140	12	2	4	11	17	4	1	1
1887571	10	183	11	3	8	14	21	5	1	1
1887571	11	72	8	1	0	6	8	2	1	1
1887571	12	95	11	1	1	7	12	3	1	1
1887571	13	172	17	2	3	10	22	3	1	1
1887571	14	209	16	0	6	12	23	5	1	1
1887571	15	130	9	1	7	8	16	5	1	1
1887571	16	646	36	3	53	27	92	8	1	1
1887571	17	777	38	7	49	43	89	10	1	1
1887571	18	137	9	1	5	9	15	2	1	1
1887571	19	34	1	0	0	6	1	7	1	1
1887571	20	12	1	0	0	2	1	1	1	1
1887571	21	238	3	0	34	8	37	3	1	1
1887571	22	85	8	0	1	7	9	2	1	1
1887571	23	375	19	1	30	14	50	6	1	1
1887571	24	22	3	0	0	2	3	1	2	2
1887571	25	64	3	1	1	5	5	2	1	1
1887571	26	184	7	1	21	8	29	3	1	1
1887571	27	105	6	0	4	5	10	2	1	1
1887571	28	91	4	0	5	5	9	2	1	1
1887571	29	37	2	0	0	5	2	2	1	1
1887571	30	59	5	1	2	5	7	2	1	1
1887571	31	60	4	0	2	5	6	2	1	1
1887571	32	60	3	0	0	7	4	2	1	1
1887571	33	93	4	0	5	5	9	2	1	1
1887571	34	42	2	0	0	4	2	4	1	1
1887571	35	396	21	1	22	26	43	10	1	1
1887571	36	123	8	2	4	12	14	5	1	1
1887571	37	674	27	0	69	30	99	6	1	1
1887571	38	90	8	1	2	8	10	4	1	1
1887571	39	127	4	2	3	6	11	5	1	1
1887571	40	123	13	0	3	8	16	2	1	1
1887571	41	286	27	1	9	24	37	9	1	1
1887571	42	80	7	1	0	6	7	3	1	1
1887571	43	35	3	0	0	5	3	2	1	1
1887571	44	297	4	0	44	9	48	0	1	1
1887571	45	24	2	0	1	3	3	0	1	1
1887571	46	378	25	3	33	19	59	5	1	1
1887571	47	69	9	2	0	5	9	3	1	1

1887571	48	304	14	0	30	8	44	4	1	1
1887571	49	35	2	1	0	5	3	3	1	1
1887571	50	41	2	1	0	5	3	4	1	1
1887571	51	38	2	1	0	5	3	3	1	1
1887571	52	34	3	0	0	4	3	2	1	1
1887571	53	36	4	0	0	4	4	2	1	1
1887571	54	69	6	0	1	9	7	4	1	1
1887571	55	46	3	0	1	7	4	3	1	1
1887571	56	33	4	0	0	4	4	2	1	1
1887571	57	23	2	0	0	4	2	2	1	1
1887571	58	7	1	0	0	1	1	0	1	1
1887571	59	23	2	0	0	4	2	2	1	1
1887571	60	73	6	0	2	9	8	4	1	1
1887571	61	38	2	1	0	5	3	3	1	1
1887571	62	14	1	0	0	2	1	2	1	1
1887571	63	28	2	0	0	5	2	3	1	1
1887571	64	41	2	1	0	5	3	3	1	1
1887571	65	32	4	0	0	4	4	2	1	1
1887571	66	80	8	1	2	3	11	1	1	1
1887571	67	38	2	1	0	5	3	3	1	1
1887571	68	54	4	0	1	4	5	1	1	1
1887571	69	85	9	0	0	6	9	4	1	1
1887571	70	51	4	0	1	3	5	1	1	1
1888010	1	82	9	2	0	8	10	2	1	1
1888010	2	145	14	2	2	15	16	7	3	2
1888010	3	333	21	2	23	15	44	6	2	2
1888010	4	112	6	1	14	2	20	0	0	0
1888010	5	93	10	1	1	6	11	3	1	1
1888010	6	180	19	2	7	14	26	3	1	1
1888010	7	611	42	4	42	35	85	14	2	4
1888010	8	148	14	1	4	10	18	3	1	1
1888010	9	669	45	5	47	30	93	16	2	4
1888010	10	344	31	4	9	19	41	8	3	2
1888010	11	489	50	3	7	33	60	9	3	4
1888010	12	183	11	4	11	16	25	6	2	2
1888010	13	176	14	2	13	10	27	2	4	1
1888010	14	59	5	0	2	4	7	2	1	1
1888010	15	260	20	0	11	17	32	6	3	3
1888010	16	35	5	2	0	5	5	1	1	1

1888010	17	41	4	0	1	4	5	2	1	1
1888010	18	102	10	0	2	8	12	4	3	2
1888010	19	145	14	0	5	14	19	3	3	2
1888010	20	225	22	4	8	24	30	10	3	4
1888010	21	162	15	2	10	15	25	7	2	3
1888010	22	43	4	0	1	3	5	2	1	1
1888010	23	235	5	0	28	9	33	2	4	1
1888010	24	35	4	0	0	4	4	1	1	1
1888010	25	247	21	0	14	14	38	3	2	2
1888010	26	941	38	2	97	29	142	3	2	2
1888010	27	301	20	2	27	17	48	4	2	2
1888010	28	189	15	0	11	11	26	4	2	2
1888010	29	185	15	0	7	14	22	8	3	3
1888010	30	260	18	0	14	10	32	4	2	2
1888010	31	403	35	2	15	25	50	10	3	5
1888010	32	127	14	2	3	13	17	5	3	3
1888010	33	249	25	2	6	18	31	9	3	5
1888010	34	47	5	1	1	4	6	2	1	1
1888010	35	97	8	1	4	8	12	2	1	1
1888010	36	85	9	0	4	4	13	0	1	1
1888010	37	116	9	0	6	7	15	1	1	1
1888010	38	40	4	0	1	4	5	2	1	1
1888010	39	500	26	1	46	21	72	6	2	3
1888010	40	86	9	1	2	9	11	5	3	2
1888010	41	183	17	1	2	14	19	7	3	2
1888010	42	311	26	0	15	15	42	4	2	2
1888010	43	152	12	0	4	11	16	7	3	2
1888010	44	42	5	1	1	5	6	2	1	1
1888010	45	412	24	2	35	21	61	5	2	2
1888010	46	42	5	1	1	5	6	2	1	1
1888010	47	92	8	1	1	8	9	6	3	2
1888010	48	536	32	4	43	22	76	12	2	4
1888010	49	136	7	0	9	6	16	6	2	2
1888010	50	69	6	1	0	5	6	3	1	1
1888010	51	109	11	3	4	9	16	4	1	1
1888010	52	378	5	0	46	11	53	4	4	1
1888010	53	190	11	1	13	14	25	8	2	3
1888010	54	129	12	1	5	12	17	2	3	2
1888010	55	627	25	1	71	24	99	2	2	2

1888010	56	77	6	0	4	7	10	0	1	1
1888010	57	843	49	1	72	30	122	11	2	5
1888010	58	262	13	0	26	13	39	1	2	2
1888010	59	259	13	0	27	13	40	2	2	2
1888010	60	226	21	3	8	17	30	7	2	3
1888010	61	31	2	0	0	2	2	3	1	1
1888010	62	97	7	0	2	9	9	4	3	3
1888010	63	370	11	0	49	12	60	6	2	2
1888010	64	406	18	0	40	19	58	8	2	3
1888010	65	263	29	8	8	26	38	8	3	5
1888010	66	719	46	5	55	35	104	13	2	5
1888010	67	94	9	0	1	7	10	4	3	2
1888010	68	210	17	1	9	15	27	4	3	2
1888010	69	93	8	0	3	6	11	1	1	1
1888010	70	74	6	0	3	5	10	0	1	1
1888011	1	57	5	0	0	8	5	6	1	1
1888011	2	344	20	1	30	16	50	9	2	2
1888011	3	684	41	0	54	24	95	10	2	3
1888011	4	194	12	0	19	4	31	1	0	0
1888011	5	647	29	0	69	17	98	5	2	2
1888011	6	1074	62	2	98	35	161	14	2	5
1888011	7	115	10	1	7	7	17	5	1	1
1888011	8	1152	75	5	74	62	151	29	2	9
1888011	9	53	5	0	1	5	6	2	1	1
1888011	10	170	18	2	2	14	20	7	3	3
1888011	11	113	10	0	3	12	13	6	3	2
1888011	12	30	4	1	0	4	4	1	1	1
1888011	13	210	22	1	7	16	29	4	2	2
1888011	14	73	8	0	1	7	9	3	3	2
1888011	15	53	4	1	2	4	7	1	1	1
1888011	16	58	7	1	0	8	7	3	3	2
1888011	17	29	3	0	0	4	3	2	1	1
1888011	18	444	27	2	34	21	63	8	2	3
1888011	19	153	4	0	19	2	23	0	0	0
1888011	20	273	17	2	21	13	39	3	4	1
1888011	21	76	7	0	5	3	12	1	4	1
1888011	22	499	23	0	53	18	76	4	2	2
1888011	23	202	18	0	9	13	27	2	4	1
1888011	24	202	9	0	16	15	26	6	2	2

1888011	25	270	18	1	21	16	39	6	2	2
1888011	26	215	5	0	27	9	32	3	4	1
1888011	27	473	25	2	41	27	66	10	2	4
1888011	28	34	4	0	0	4	4	2	1	1
1888011	29	82	8	1	0	12	8	6	3	3
1888011	30	27	2	0	0	4	2	3	1	1
1888011	31	313	10	1	33	15	45	4	2	2
1888011	32	42	5	0	0	5	5	3	1	1
1888011	33	86	8	0	4	8	12	4	1	1
1888011	34	118	10	0	6	8	16	2	4	1
1888011	35	184	20	2	2	21	22	9	3	4
1888011	36	61	6	0	0	8	6	4	3	2
1888011	37	479	18	0	59	17	77	4	4	1
1888011	38	400	28	1	29	31	58	8	2	3
1888011	39	318	17	0	36	13	53	2	2	2
1888011	40	55	5	0	1	5	6	2	1	1
1888011	41	29	3	0	0	4	3	2	1	1
1888011	42	27	2	0	0	5	2	3	1	1
1888011	43	34	4	0	0	4	4	2	1	1
1888011	44	35	3	0	0	4	3	2	1	1
1888011	45	92	9	0	1	10	10	5	3	2
1888011	46	33	4	0	0	4	4	2	1	1
1888011	47	35	3	0	0	4	3	2	1	1
1888011	48	29	3	0	0	4	3	2	1	1
1888011	49	679	45	1	56	49	103	11	2	6
1888011	50	36	3	0	0	6	3	4	1	1
1888011	51	402	9	0	45	18	56	5	4	1
1888011	52	22	2	0	0	4	2	2	1	1
1888011	53	146	11	1	11	10	22	2	4	1
1888011	54	189	13	0	15	6	28	1	4	1
1888011	55	49	5	0	0	7	5	4	3	2
1888011	56	32	3	0	0	5	3	3	1	1
1888011	57	23	2	0	0	4	2	2	1	1
1888011	58	390	37	3	13	32	50	13	2	5
1888011	59	198	16	1	11	13	28	5	2	2
1888011	60	279	26	0	11	18	38	5	2	4
1888011	61	28	3	0	0	4	3	2	1	1