Master's Project Proposal

"ECfAIR - Email Classification for Archival and Institutional Repositories: A Semi-Supervised Classification Method for Email Collections"

Introduction

I hope to explore text mining applications for processing large email collections as part of a digital archives workflow. As archival materials become more and more digital, email corpora have become essential components of many born-digital collections. Archivists must contend with thousands and thousands of individual documents upon accepting a collection. Given the time and infrastructure requirements associated with adequately processing these large bodies of information, many archivists cannot perform crucial appraisal steps on these kinds of collections. In other words, the means for manually sorting these materials and deciding what to keep and how to organize it are beyond the resources for many institutions.

Text mining in particular and machine learning more generally show much potential for surmounting these obstacles. If we could utilize algorithms to build models that could reliably classify emails according to criteria provided by archivists, we could significantly decrease the resource requirements for adequate appraisal and processing of digital collections. In short, I hope to explore methods and tools that could potentially be used to quickly classify large bodies of email documents given minimal training data. Successful completion of this project will provide archivists with a method for processing large digital collections without extravagant overhead.

My project will involve the development and evaluation of a tool that will automatically categorize email messages based on their text contents and provide confidence rankings for the classification predictions. This tool will serve as a proof of concept for the archival application of a semi-supervised classification method sometimes referred to as "predictive coding." In short, this tool will accomplish the following task:

1. I will provide a large corpus of unlabeled emails accompanied by a small "seed" set of labeled (pre-categorized) emails.
2. I will designate one category ("target class") into which the seed articles have been sorted and provide a desired recall threshold.
3. The software will use the seed set to "learn" the collection and output a list of all the original documents, each with a rank value of 0-9 indicating how confident the software is that the document should be categorized into the target class.

If successful, this tool can help archivists gain a better understanding of their collections' contents at an item-by-item level. While I do not envision the product being very useful as a standalone tool, it can serve as a starting point for researchers hoping to build upon the classification method described above. It can also be used by developers hoping to expand the capabilities of existing archival tools for born-digital processing and appraisal. If my evaluation of the tool shows that my proposed method was not successful, the lessons learned will still be useful from the above perspectives.

Literature Review

Born-Digital Records: Problems and Solutions

The archival community has recognized the need to introduce automation to their workflows for several decades.[1] The volume of born-digital records flowing into archival institutions has presented archivists with a wide array of challenges. How best can we store said materials? Perhaps more importantly, how can we adequately sort and appraise collections that regularly contain hundreds of thousands or even millions of individual records? As one researcher describes it, institutions dealing with an "overwhelming volume of digital information" must find ways to "separate out the good oil of meaningful records from vast quantities of information sludge…Put simply, without appropriate attention (and suitable technological aids), we will drown in the sludge."[2]

Faced with such challenges, many in the field have recognized that traditional methods for dealing with archival collections will no longer prove adequate for many incoming and existing collections. According to the National Archives of the UK, "There is a general acceptance that processes designed for the review of paper records collections will not meet the challenges of born-digital records."[3] Such issues have led many archivists and researchers to call for an increased focus on automated processes and workflows. For example, the necessity of an "automatic appraisal mechanism" to help archivists properly manage a deluge of born-digital records has been recognized at least since the mid 1990s.[4]

A wide range of tools and applications have since surfaced to help mitigate some of the issues related to born-digital materials. Open-source environments and packages such as ArchivesSpace[5], Archivematica[6], and BitCurator[7] feature a wide range of tools to aid in tasks like ingest, description, and preservation. For institutions with significant financial resources, proprietary platforms like Preservica[8] are also an option. ArchivesSpace, to provide one example, supports "core functions in archives administration such as accessioning; description and arrangement of processed materials including analog, hybrid, and born-digital content; management of authorities (agents and subjects) and rights; and reference service."[9] Similarly, the BitCurator environment provides users with a means for accomplishing tasks like "pre-imaging data triage, forensic disk imaging, file system analysis and reporting, identification of private and individually identifying information, and export of technical and other metadata."[10] Ingest, processing, and long-term preservation & storage all represent important steps in any archival workflow. These steps are well-covered by tools like those mentioned above. However, the important step of appraisal, which must precede all other steps in an archival workflow and continues to occur throughout a collection's lifecycle, continues to present obstacles for those hoping to more fully automate archival workflows.

Born-Digital Records: An Appraisal Challenge

Appraisal may be defined as "any selection activity that enables archivists to identify recorded information that has enduring value, primarily for the documentation of modern society."[11] More specifically, we can think of appraisal as the:

> *Process of determining the value and thus the disposition of records based upon their current administrative, legal, and fiscal use; their evidential and informational value; their arrangement and condition, their intrinsic value; and their relationship to other records.*[12]

This task, though essential to any collection, presents unique challenges when dealing with electronic records. For one, "The time taken to render and to read even a representative sample of these large volumes [of information] is problematic."[13] It is simply not feasible to sort through many collections at the level necessary to ascertain their contents and worth. Indeed, this has led some institutions just to store everything "in the hope that [sorting methods] may eventually be found" to separate the important from the unimportant.[14] This solution leads to its own issues, however, for "even if the resources and the technical ability are available, preserving everything is probably not desirable. Appraisal cuts both ways; it is just as important to determine what to exclude from collections as it is to determine what to include."[15] Such considerations make automation an attractive option for many institutions pondering the appraisal of born-digital materials. Indeed, the researchers from the National Archives of the UK predict that "Future appraisal, selection and sensitivity review of digital records will require the assistance of technology."[16]

A major issue with automating the appraisal process is that it is highly context-specific.[17] What is important about one collection to one institution may be entirely different from the important aspects of the same collection in other institutions. Similarly, each collection within an institution will probably require different appraisal considerations than those of other collections. While automation can provide solutions to problems of scale, a combination of scale and context-dependency make appraisal a difficult challenge. In short, there is no one-size fits all solution for the appraisal of born-digital materials.

Email: A Valuable Record

Email collections constitute an important subgroup of born-digital records. As one digital curation manual describes it, email "is the modern day equivalent of paper-based correspondence" and has thus assumed significant "cultural, historical, or research" value.[18] Others have since echoed this sentiment, arguing that email "is the natural successor to the written letter, allowing for snippets of "unformed thought and candor that only the recipients are both privileged and sensitive enough to receive." As a result, the same author argues, "Email collections have the potential to capture the emergence of thoughts and decisions that illustrate how and why things developed the way they did."[19] Because of their central role in modern communications, it is no wonder that institutions are exploring ways to best curate their email collections.

Email: Appraisal Challenges

Accomplishing this goal, especially in regards to appraisal, brings numerous challenges. "The volume and unstructured nature of email makes its management, disposal and sensitivity review difficult," a recent study found. In one of their case studies, for example, the same

researchers found that "over 20 years of routine backups have resulted in an unwieldy backlog amounting to 67,000 tapes and 28 petabytes of content."[20] Problems of scale have led others to discern that, "the traditional practice of reviewing records at an item-by-item level for restrictions may need to be reconsidered as the growth of records increases."[21]

The complex and multifaceted nature of the records themselves also makes management difficult. As Prom (2011) makes clear, "a single email account contains records of disparate context, structure and content, documenting activities both mundane and extraordinary."[22] Individual emails themselves come with a wide range of attachments and embedded references and are deeply bound to their threads/context and the technical systems from which they originated. As the aforementioned digital curation manual explains, "Email curation is a many-layered thing."[23]

Another concern is privacy. Email collections can and often do contain sensitive information that must be protected. Cocciolo (2016) argues that "making email collections available in publicly accessible archives faces a number of challenges, including ensuring that personal privacy is preserved and the institutions making these collections available do not inadvertently expose themselves to liability."[24]

Combined with the challenges of providing adequate storage for and access to collections with the volume that usually characterizes email collections, privacy concerns make the need for adequate appraisal all the more pressing. According to Cociollo (2016), "Email appraisal is essential to both conserve IT resources and protect personal privacy." Similarly, another researcher argues, "Identifying various types of restrictions that must be placed on email collections prior to making them publicly available is a notable challenge at scale as is how to easily reduce the number of non-archival messages."[25]

Email: Existing Appraisal Solutions

In the past decade, a number of organizations have introduced tools to aid archives in the curation of their email collections. ePADD, a software package developed by Stanford Libraries, "supports the appraisal, processing, preservation, discovery, and delivery of historical email archives."[26] The appraisal module allows users to browse messages, review attachments and folders, and attach annotations and labels to email. Notably, the module also includes entity extraction and lexicon analysis tools to give users an automated means for identifying the contents of large numbers of individual emails.[27]

Another recent project called TOMES (Transforming Online Mail with Embedded Semantics) has developed a tool that converts emails to an easy-access, standardized format and performs entity-extraction to aid in the sorting of government-related emails.[28] This tool, currently under development by a team from the state archives of North Carolina, Kansas, and Utah, can help archives hoping to appraise their email collections by making email files more transparent and by automatically identifying certain topics contained within. By doing so, institutions can more easily work with and preserve their files and better prioritize certain sections of a given collection over others.

A team at the University of North Carolina's School of Library and Information Science is also working to develop solutions to some of the aforementioned problems through a project called RATOM (Review, Appraisal, and Triage Of Mail).[29] The RATOM Project seeks to augment

the existing capabilities of both the TOMES software and the BitCurator digital preservation environment. Among other things, the project's deliverables will "explore the use of machine learning to separate irrelevant emails from those that should be preserved, and will apply natural language processing methods to identify topics of interest within those records so the messages can be tagged for improved organization and retrieval."[30] These functions will be particularly useful for tackling the problems of scale mentioned above and will help streamline the identification and proper treatment of sensitive content contained within email collections.

Predictive Coding: A Novel Solution for Automatic Binning of Email Collections

All of the aforementioned tools serve to dramatically expand an institution's toolkit for working with email collections. Automatic identification of sensitive materials, file-format conversion, topic extraction, browsing capabilities, and integration with other software packages all serve important purposes. That being said, there exist other methods for applying automation more generally and machine learning in particular to email collections that as of yet have not been fully explored by the archival community. One technique in particular, known as "Predictive Coding" in the legal community, holds much promise for approaching the appraisal challenges presented by large email collections.

According to the National Archives of the UK, predictive coding:

> *"…is a way of automatically classifying documents based on statistical analysis and machine learning…It involves a learning process, which requires the reviewer to identify a relevant subset of information from a larger collection to train the software. Algorithms in the software then use this 'seed set' to find conceptually similar information in the larger collection."*[31]

In other words, a user (in our case, an archivist) can manually label or categorize a small subset of a larger collection (in our case, a collection of emails), which can then be used by machine learning algorithms to classify the rest of the collections according to the categories set out by the archivist. This could prove useful in the appraisal process, for it would allow archivists, alone or in concert with donors or subject experts, to label small subsets of collection and extrapolate those labels to the rest of the collection. Not only would this aid in the future arrangement of materials the archive does decide to keep, but it will also help institutions sort what they should keep and what they should not keep.

This technique has been put to use by the legal community over the past decade to help attorneys sort through large volumes of case documents. In the words of one information professional, "It is difficult and time-consuming to sift through e-mails, texts, contracts, spreadsheets, and other types of ever-increasing media. While subject matter experts are still necessary, predictive coding eliminates much of the manual work and time from the task." A legal professional exploring the technology finds that, "The greatest advantage of predictive coding is the potential to dramatically reduce the number of documents requiring attorney review."[32]

Members of the archival community have already begun to explore ways for applying this technique to their workflows. One project in particular, a joint effort between the Illinois

State Archives and the University of Illinois called the Capstone Email Project, seeks to do just that. In the words of the team members, the Capstone project will first secure access to a large email collection, then "explore tools that use technology-assisted review techniques (predictive coding in particular) for the purposes of parsing and classifying the email."[33] As of 2018, the project had identified a number of proprietary (Microsoft's 365 Advanced eDiscovery,[34] Luminoso,[35] Axcelerate (Recommind)[36], and Ringtail[37]) and open-source (ePADD, TAR Evaluation Toolkit[38]) software packages for evaluation purposes. Out of all of these, the researchers found that Ringtail offered the most useful functionality and adaptive interface for archival applications related to email.[39]

Predictive Coding: The Need for Open-source Alternatives

Though the final results of the Capstone project have yet to be published, we can identify further avenues for exploration from their progress so far. Notably, neither of the open-source packages they evaluated provided the functionality necessary to apply predictive coding to email collections. While the proprietary platforms seemed more promising, these solutions by nature lack transparency that can allow other developers (such as those working to integrate a number of existing tools like TOMES and RATOM) to build on their work. From the archivist's perspective, this lack of transparency could also prove problematic if an institution is ever asked to justify choices made based on the software's output. As such, it stands to reason that there is room for more free and open-source development of predictive coding solutions targeted towards the appraisal of email collections. To prototype a new method for predictive-coding that can be openly accessed and shared among institutions is just what my project seeks to accomplish.

Machine Learning: Definitions

Before describing the particular method I hope to adopt, I will provide a short description of existing machine-learning (ML) methods and their application to email collections. To begin, most ML methods fall into one of two categories: "unsupervised" methods that create categories from scratch and group documents without any input from the user and "supervised" methods that require large sets of pre-labeled documents to "learn" a collection and build a classification model. Supervised methods suffer from the significant drawback that adequately sized training sets require a significant amount of labor and are often prohibitively expensive to produce. On the other hand, unsupervised techniques may produce categorization schemes that have little to no bearing on the realities of the collection or the needs of the user.

Machine Learning: Applications for Email Collections

Many of the attempts to utilize ML to help manage email collections have fallen into one of these two categories. Topic extraction tools, such as those included in the ePADD or TOMES packages, fall into the unsupervised category. Essentially, these tools use algorithms to discern patterns and themes within large datasets and to generate topic tags that can be used to

describe said patterns. Similarly, others have used unsupervised methods to automatically generate ontologies for large email collections.[40] In other words, these systems automatically recognize and sort documents into *algorithm-derived* categories, after which and archivist can then decide whether or not those categories are useful and/or adequate for their purposes. Such tools are useful in that they can give an archivist an idea of the contents of the collection and of the associations between particular documents or subsets of documents. On the other hand, they work entirely independent from user input and prevent archivists from having any say in how their records are grouped.

Others have tested "supervised" methods for use in archives in general and with email collections more specifically.[41] As mentioned above, supervised methods suffer from the drawback that they require large training sets for the algorithm to "learn" a collection. This is a drawback because it forces archivists to spend the time and effort required to label enough data for ML algorithm to produce a workable classification model. This can defeat the purpose of using such technology entirely. As Roland, et al. explain, "The cost and time needed to configure machine-learning solutions – in particular, developing large, clean and labelled training and testing sets of data" presents a significant obstacle to many institutions. [42]

Another consideration is that models derived from supervised methods are extremely context-specific. A model trained using one dataset oftentimes cannot be transferred for use with another collection. One study testing Support Vector Machines (SVMs), a popular and proven type of supervised algorithm, for use with email found that, "the poor, near random precision and near-zero Kappa coefficients of each SVM when used to classify "business value" in [the group of emails not used for training the model] indicates that each SVM is very context-sensitive…an e-mail of "business value" that deserves to be kept as a record in one context can be entirely irrelevant in another."[43]

<u>Predictive Coding: Bridging the Gap</u>

Predictive-coding can be described as a "semi-supervised" ML method. As such, it falls somewhere between the two major categories above and helps mitigate the drawbacks of both. The application of this method to archival work with email collections will help fill a role to which many existing systems are not well-adapted. Why does predictive coding represent such a departure from the supervised and unsupervised methods described above?

First of all, by requiring that the archivist only present a small number (anywhere from dozens to hundreds) of "seed" documents as opposed to thousands of training documents, this technique eliminates much of the costs associated with traditional supervised methods. Predictive coding also provides an advantage over the topic extraction methods mentioned above because it works using categories supplied by the archivist rather than the other way around. While the usefulness of topic extraction tools cannot be denied, this "semi-supervised" system allows the archivist to provide some degree of human guidance without asking for too much of it. Finally, the specific method I hope to implement (see below) ranks documents according to a confidence rating as opposed to simply placing them in one bin or another. This feature gives users a higher degree of agency by allowing them a convenient means for evaluating how well individual documents might fit into their provided categories.

Summary: Tying It All Together

To summarize, electronic records and email in particular have gained considerable importance for archival institutions. Email records represent important pieces of the historical record, yet many institutions have come to realize that existing archival workflows and tools do not meet the needs of such collections. Appraisal of email in particular has become an object of focus, for archives cannot and should not attempt to preserve the entirety of all incoming email collections. That being said, the volume of these collections as well as their technical complexity and unstructured nature present significant appraisal challenges.

Many ongoing research projects and existing software packages provide functionalities for tackling the problems of email appraisal. Features like automatic identification of sensitive content, topic extraction, format-conversion, and browsing/annotation interfaces make the task of appraisal easier for archivists in many ways. Predictive coding will prove a useful addition to existing capabilities by allowing archivists or other subject experts to label just a few emails in a collection and then extrapolate the resulting categories to the rest of the collection. At least one project (the Capstone Email Project) has begun to explore predictive coding as it relates to email in archives, but as of this writing there are few examples of open-sourced software targeted towards this specific use.

My project will result in a tool that adapts a particular predictive coding method from a different domain to use with email collections. I intend the tool to serve as proof of concept for this particular method and will evaluate it to get a sense of how successful it might be when put to this particular use. While I do not expect the tool to serve as a production-ready standalone piece of software, it is my hope that others can use it as a jumping-off point for inclusion in larger, more multi-faceted software packages. I will now describe the architecture I hope to implement as well as the methods I will use to evaluate it.

Research Question

1. How effectively can a tool based on the semi-supervised classification system developed by Varghese, et. al (2018) classify email messages from a real-life corpus based on the textual components of the ‘subject’ and ‘body’ fields of each message?

Methods

Conceptual Model

The tool I plan to build will serve as proof of concept for the archival application of a novel classification method for largely unlabeled collections. This method, a type of “ensemble learning,” was originally developed by research contractors at the US EPA for the automatic classification of research articles during large literature searches. Essentially, this particular method asks a user to supply the system with a small number of labeled “seed” documents from a large collection of otherwise unlabeled (un-categorized) materials. The system uses the seed articles to “learn” the collection and rank the remaining documents on a confidence scale

to indicate whether or not each document might be associated with categories supplied by the user.[44]

This system, developed by Varghese, et. al., has several characteristics that make it well suited for archival applications. To begin, it is useful to sketch a general outline of the architecture they present in the aforementioned study:

1. A small subset of a large collection of documents is labeled according to one or more categories. These are the 'seed' documents.
2. The researchers set a specific recall value they would like to meet. In this case, recall refers to the proportion of all relevant documents that are predicted to be 'relevant' by the system. In their study, Varghese, et. al. use 95% as a desired recall threshold.
3. The entire set, including the seeds, are clustered using three separate algorithms: K-Means, Latent Dirichlet allocation (LDA), and Non-negative matrix factorization (NMF). Each algorithm is run 3 times, with values for k (the number of output clusters) set to 10, 20, and 30, respectively.
4. For each run of each algorithm, the distribution of seed documents across the resulting clusters is determined (i.e. 10 in one cluster, 7 in another, 2 in another, etc.).The system then determines the minimum number of clusters needed to meet the desired recall value in regards to the total number of 'relevant' seed articles. This provides our classification scheme. All documents that lie within that minimum set of 'relevant' clusters are then classified as 'relevant.'
    a. To provide an example, imagine a scenario in which k=10 (the number of clusters), the desired recall value r=.7, and the seed set contained 50 'relevant' articles. After running our algorithm, we find that 18 'relevant' seed articles lie in one cluster, 11 lie in another, and 6 lie in another. Together, these 3 clusters account for 70% of all seed articles, which meets our desired recall threshold. As a result, we classify every document, seed and non-seed, within those three clusters as 'relevant.' All documents that did not fall into these 3 clusters are classified as 'non-relevant.'
    b. Note – to expand this idea to include multiple algorithms working in concert, recall thresholds for individual models are scaled up iteratively from 50% to 90% until the overall recall threshold (in their case, 95%) is met.
5. To classify documents using the combined "ensemble," method, any document classified as 'relevant' by at least one algorithm is classified as relevant overall. Priority scores are also calculated for each document. Each document labeled as 'relevant' in each pass of each algorithm receives one point towards its priority score. Each document in the original set can thus receive a total score of 9.
6. The output of this system consists of binary classifications (relevant/non-relevant) and 'priority' scores (0-9) for each document. Traditional evaluation metrics like precision, recall, and F1-score can be used to evaluate performance. The authors also add 'elimination rate', which is defined as the "proportion of the initial set of documents that is eliminated from review using the automated process", to their set of evaluation metrics.[45]

Archival Applicability

As mentioned above, this system holds potential for archival applications for a number of reasons. For one, "the cost function used in this method to determine cluster labeling is entirely based on achieving the recall requirement."[46] Recall is important in an archival context because while it is useful to eliminate documents that should not be considered for preservation, it is even more important to ensure that all documents of worth are preserved. In other words, focusing on recall results in a system that prioritizes keeping relevant documents over disposing of irrelevant documents.

That said, there would be little point in pursuing such a system if it did not eliminate some significant portion of a collection from being considered as 'relevant.' It bears repeating that, "The greatest advantage of predictive coding is the potential to dramatically reduce the number of documents requiring…review."[47] If a system did not dramatically reduce the processing and storage requirements for digital collections, it would be of little use as an appraisal aid. It is with similar considerations that Varghese, et. al. included "elimination rate" as a metric in their evaluation. It is true that the proper elimination rate changes from collection to collection given the specific contents of each, but including this metric can give us an idea of how useful a successful classifier can be.

How much effort might the use of such a system save an archival institution if one were to achieve similar results to those of Varghese et. al.? The results are promising. According to the authors, the above system achieved the following results using just 25 'relevant' and 25 'non-relevant seeds' out of a collection of 6800 documents: recall = .91; F-1 score = .81; elimination rate = .63. Translated into plain language, this means that their system was able to successfully identify 91% of all 'relevant' documents while eliminating 63% of the total collection from contention as being potentially relevant. From an archival perspective, this would mean that an archivist could potentially identify 91% of all emails belonging to a certain category while also eliminating the possibility that over half of the documents in a 6800-email collection might belong to that category. If these results do hold true, they could be achieved simply by hand-labeling a set of 50 seed documents from the original collection. Such a system would prove useful indeed.

Description of Tool

The tool will consist entirely of Python scripts that can be installed and run via the command line. I plan to build it using the latest stable version of the popular Anaconda distribution, which includes many of the necessary packages. By using Anaconda, I can allow for the quick download and setup of my tool across platforms without a long list of external dependency requirements. For much of the data processing and machine learning component I plan to use the popular and well-established Sci-Kit Learn package. In order to meet the computational requirements of such a large dataset, I hope to make use of one or several computing resources at UNC, such as the Virtual Computing Lab[48] or the university's Research Computing[49] services. To this end I have spoken with staff at UNC Libraries and at SILS and confirmed that such resources will be available. I plan to fully document my code using a Github repository with source code and Jupyter notebooks describing individual processes.

The workflow I will implement goes as follows:

1. I will select a .csv file containing a test collection of emails. The data will be structured using columns for email subject, email body, a column each for all categories into which my data has been sorted. In this case, I will use existing categories to label all documents as being "business-related" or "non-business-related." The data will also include a category indicating whether or not each document belongs to the seed set.
2. The system will then vectorize the documents, transforming each into its TFIDF (Term Frequency-Inverse Document Frequency) representation. To reduce computational requirements, I will employ sparse matrices following TFIDF transformation (see Varghese, et. al., page 402). Again, my seed articles will be structured to include labels for "business-related" or "non-business-related".
3. If time permits, I will develop a subsystem that will inform me as to whether or not the clustering algorithms chosen will be effective for the task at hand. This will give me an idea as to how successful I can reasonably expect the classification task to be. Including this subsystem will improve the reusability of my tool for other researchers that may want to adapt my code for use with their own collections.
    a. To test an individual clustering algorithm against their set of seed articles, one can issue a command to run that algorithm over their set of seed articles iteratively with all values of k from k=1 to k=n.
    b. For each iteration, all clusters (which do not necessarily contain the same number of documents) will be labeled according to the user-supplied target variable value for the selected seed article category. In other words, if a user hopes to test according to their supplied category of "business-related" emails, the test will label each cluster if a majority of the documents within have been labeled as "business-related" by the user.
    c. Once all clusters have been labeled, accuracy for each will be computed. If a cluster has 2 "business-related emails" and 1 non-business email, that cluster will receive an accuracy score of 67%. The average accuracy rating across all clusters for that value of k will be computed.
    d. This process is repeated for all values of k, allowing the system to plot the algorithm's accuracy scores along a curve, with k being the horizontal axis and average accuracy being the vertical axis. A curve that shows a quick accuracy jump as k begins to increase will indicate that the algorithm in question is capable of producing meaningful groupings for the data set in question.
    e. This process can be repeated for all available clustering algorithms so that I can pick and choose which algorithms for which values of k to use for the following steps.
4. After evaluating each algorithm against their seed articles, the system will ask me to select up to 3 clustering algorithms with three k values for each. I will build my tool to include the following three algorithms because these were used by Varghese, et. al. to promising results in their original study: k-Means, Nonnegative Matrix Factorization (NMF), and Latent Dirichlet allocation (LDA).[50] More algorithms can be added if time/scope permits. The system will then ask me to designate a minimum desired recall

threshold in regards to the classification category in question ("business-related," etc.). To maintain consistency with the original study, I will start by using .95 as my desired recall threshold

5. After setting these parameters, the system will run the chosen clustering algorithms for each value of k over the entire data set using the TFIDF feature matrices. For each algorithm / k-value combination, documents will be classified as either belonging or not belonging to the target variable category based on the category-positive seed articles in their specific cluster. For each run, the system will assemble the minimum number of clusters required to account for the desired recall rate of seed articles and label all articles in those clusters as belonging to the target variable category. For example, if we desire a recall rate of .95, we could pull together the minimum number of clusters required to account for 95% of relevant documents and label all items in those clusters as "relevant." While the overall recall rate will be set by the user, the system will determine the recall threshold for the individual algorithms required to meet the overall threshold. As did Varghese, et. al, I will do so iteratively checking individual thresholds of .5 to .9.
6. If a document is labeled by at least one algorithm as being relevant, it is classified as 'relevant' according to the overall classification scheme. For each run of an individual algorithm that identifies a document as 'relevant,' +1 will be added to its 'relevance score'. Documents can thus receive scores (or ranks) ranging from 0-9 if I end up using three algorithms with three k values each
7. The system will return the original .csv file with columns for predicted classifications and rankings.
8. I will then perform evaluation on the system using methodologies described below.

Evaluation: Traditional Metrics

I plan to build and test my tool using the Enron Email dataset that was released to the public by the Federal Regulatory Commission in 2003.[51] Fortunately, I have located a version on Data.World[52] that has been partially labeled (~1700 emails) by graduate students at UC Berkley.[53] This version of the dataset has been converted from the original PST files to .csv format with columns for body, subject, etc. as well as for all of the label categories. I will use these pre-labeled documents as my seed articles by grouping them according to their "business-related" or "non-business-related" status.

In order to evaluate my system, I will hand-label a randomly selected, 1000-instance test set in accordance with the target categories mentioned above. This set will remain within the overall corpus as the system performs its analysis. During this process the labels will be ignored by the system. Upon completion of the classification/ranking task, I will use the classifications I assigned by hand to compare results with the system output. I will compile actual recall, actual precision, elimination rate, and F-1 score to gain an idea of the system's success. I will also determine and analyze the rankings distribution across the test set to gain an idea of how useful such a system can be. Does the precision increase or decrease as one goes down in rankings (from 9-0)? As the creators of the original system explain, one would expect fewer false positives amongst the higher ranks if the system performs as expected.[54]

Evaluation: Limitations and Future Prospects

Evaluating performance based on these traditional metrics will give me some idea of how well my system performs its task. Unfortunately, however, the context-specificity of appraisal tasks means that any results I might gain with my test collection may not hold if another researcher were to run the same tests using a different email collection. Thus, the more fundamental purpose of this project is to give others a framework by which they can test this or similar methods using their own datasets. By making my adapted version of this particular ensemble method available to others, I will allow archivists and researchers to potentially test the method with real-world collections to which I do not have access. In other words, my purpose is not solely to test and evaluate results for a single data set, but also to build a framework by which others might do so with their own data. Only through a wide range of tests and evaluations can one make any conclusions about the general efficacy of this method in regards to appraising email collections. Due to limits on scope and data accessibility, this project cannot and does not promise to make such conclusions. It can, however, provide foundations on which others can build and work towards more comprehensive investigations.

Proposed Project Timeline

- March 2019: Submit Proposal, begin development.
- End of June 2019: Complete working version of tool on my local machine using a smaller subset of the Enron Email Set.
- End of August 2019: Complete trial runs using the entire Enron corpus with extended computational resources. Make sure that the tool is fully portable to virtual machines across platforms and environments. Ensure that all documentation and source code is available on GitHub.
- End of September 2019: Complete first draft of the written component.
- Middle of November 2019: Submit written component and online materials to advisor for grading.

References

[1] Gilliland-Swetland, Anne Jervois. “Development of an expert assistant for archival appraisal of electronic communications: An exploratory study.” Doctoral Dissertation. The University of Michigan, 1995. Available from ProQuest Dissertations Publishing (9542845). Retrieved from http://libproxy.lib.unc.edu/login?url=https://search.proquest.com/docview/304205621?accountid=14244
[2] Rolan, Gregory, et. al. “More human than human? Artificial intelligence in the archive.” *Archives & Manuscripts* 47, no. 2 (2019): 180. Retrieved from https://www.tandfonline.com/doi/full/10.1080/01576895.2018.1502088
[3] “The Application of Technology-Assisted Review to Born-Digital Records Transfer, Inquiries and Beyond." National Archives of the UK (2016): 7. Retrieved from https://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf
[4] Gilliland-Swetland 1995, 20.
[5] https://archivesspace.org/
[6] https://www.archivematica.org/en/
[7] https://bitcurator.net/
[8] https://preservica.com/about
[9] https://archivesspace.org/application/demo
[10] https://bitcurator.net/bitcurator/
[11] Cox, Richard J., and Helen W. Samuels. "The Archivist's First Responsibility: A Research Agenda to Improve the Identification and Retention of Records of Enduring Value." *The American Archivist* 51, no. 1/2 (1988): 28-42. Retrieved from www.jstor.org/stable/40293193
[12] Gilliland-Swetland 1995, 3.
[13] Harvey, Ross and Thompson, Dave. “Automating the appraisal of digital materials.” *Library Hi-Tech* 28, no. 2 (2010): 314. Retrieved From https://doi.org/10.1108/07378831011047703
[14] Rolan, et. al 2019, 180.
[15] Harvey and Thompson 2010, 315.
[16] National Archives of the UK 2016, 26.
[17] Harvey and Thompson 2010, 316
[18] Pennock, Maureen. “Curating E-mails: A life-cycle approach to the management and preservation of e-mail messages.” Digital Curation Centre: Digital Curation Manual (2006): 14. Eds S. Ross and M. Day. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/curating-e-mails/curating-e-mails.pdf
[19] Cociollo, Anthony. “Email as cultural heritage resource: appraisal solutions from an art museum context.” Records Management Journal 26, no. 1 (2016): 69. Retrieved from https://www-emerald-com.libproxy.lib.unc.edu/insight/content/doi/10.1108/RMJ-04-2015-0014/full/html
[20] Rolan, et. al 2019, 188
[21] Vinh-Doyle, William P. “Appraising email (using digital forensics): techniques and challenges.” *Archives & Manuscripts* 45, no. 1 (2017): 27. Retrieved from https://www.tandfonline.com/doi/full/10.1080/01576895.2016.1270838
[22] Prom, Christopher J. “Preserving Email - DPC Technology Watch Report 11-01.” Digital Preservation Coalition, December 1, 2011. pp10-11. Retrieved From http://www.dpconline.org/component/docman/doc_download/739-dpctw11-01pdf
[23] Pennock 2006, 37.
[24] Cociollo 2016, 69.
[25] Kaczmarek, Joanne and West, Brent. “Email Preservation at Scale: Preliminary Findings Supporting the Use of Predictive Coding.” Capstone Email Project Preliminary Report, The University of Illinois (2018): 1. Retrieved from https://uofi.app.box.com/s/wxayre862oyik8c4fdx2zvcz4putk06l
[26] https://library.stanford.edu/projects/epadd
[27] “ePADD Installation and User Guide, Version 7.2”. Stanford University Libraries (February 2020): 18-57. Retrieved from https://docs.google.com/document/d/1CVIpWK5FNs5KWVHgvtWTa7u0tZjUrFrBHq6_6ZJVfEA/edit

[28] Tyndall-Watson, Camille. “April 2018 Report.” North Carolina Department of Natural and Cultural Resources: TOMES Project Report (April 2018): 2. Retrieved fromhttps://files.nc.gov/ncdcr/TOMES/20180430_nar5005__tomes_report_FINAL.pdf
[29] http://ratom.web.unc.edu/
[30] https://ischoolsinc.org/blog/category/university-of-north-carolina-unc/
[31] National Archives of the UK 2016, 8.
[32] Hampton, Wallis M. “Predictive Coding: It’s Here to Stay.” *Practical Law* (June/July 2014): 32. Retrieved from **https://files.skadden.com/sites%2Fdefault%2Ffiles%2Fpublications%2Flit_junejuly14_ediscoverybulletin.pdf**
[33]https://www.aits.uillinois.edu/services/professional_services/rims/about_rims/projects/processing_capstone_email_using_predictive_coding/
[34] https://docs.microsoft.com/en-us/microsoft-365/compliance/overview-ediscovery-20
[35] https://www.basistech.com/partner/luminoso/
[36] https://www.opentext.com/products-and-solutions/products/discovery/axcelerate
[37] https://www.ringtail.com/
[38] https://cormack.uwaterloo.ca/tar-toolkit/
[39] Kaczmarek and West 2018, 2-3.
[40] Yang, Hui and Callan, Jamie. “Ontology generation for large email collections.” The Proceedings of the 9th Annual International Digital Government Research Conference (May 2008): 259. Retrieved from https://dl.acm.org/doi/abs/10.5555/1367832.1367875259
[41] Li, Min, et al. “Business Email Classification Using Incremental Subspace Learning.” The Proceedings of the 21st International Conference on Pattern Recognition (November 11-15, 2012): 625-626. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6460212; Rolan, et al. 2019, 191-193.
[42] Rolan, et al. 2019, 186
[43] Alberts, Inge and Vellino, Andre. “The importance of context in the automatic classification of email as records of business value: A pilot study.” Proceedings of the American Society for Information Science and Technology 50, no. 1 (May 2014): 2. Retrieved from https://doi.org/10.1002/meet.14505001112
[44] Varghese, Arun, Cawley, Michele, and Hong, Tao. “Supervised clustering for automated document classification and prioritization: a case study using toxicological abstracts.” Environment Systems and Decisions 38 (2018): 398-414. Retrieved from https://doi.org/10.1007/s10669-017-9670-5
[45] Ibid., 401-405.
[46] Ibid., 402.
[47] Hampton, 2014, 32
[48] https://vcl.unc.edu/index.php?mode=selectauth
[49] https://its.unc.edu/research-computing/
[50] Varghese, et. al., 2018, 402.
[51]Leber, Jessica (2013, July). The Immortal Life of the Enron Emails. Retrieved from https://www.technologyreview.com/s/515801/the-immortal-life-of-the-enron-e-mails/
[52] https://data.world/brianray/enron-email-dataset
[53] http://bailando.sims.berkeley.edu/enron_email.html
[54] Varghese, et. al., 2018, 411.