Computational Thinking in Archival Science Research

William Underwood

Developing a Computational Framework for Library and Archival Education Workshop
iSchool Conference, College Park, Maryland, April 3-4, 2019
Overview

- Motivation for Integrating Computational Thinking into UMD MLIS program in Library and Archival Studies.
- Examples of Computational Thinking Practices being used in Archival Studies Research
Motivation for Introducing Computational Thinking into Library and Archival Studies Curriculum

- A basic understanding of the characteristics of these digital materials is important for future archivists.
- Archival collections are increasingly composed of digital materials.
- The tools and practices associated archival activities are increasingly dependent on computing.
- The way users interact with archival collections reflects the increasingly computationally-mediated nature of our world.
- For today’s learners to succeed in future archival tasks, it is essential that computational thinking is included as part of their training.
Archival Studies Research Projects Incorporating Computational Thinking Practices


The records of the WRA (Record Group 210 from 1941-47) at the National Archives in Washington D.C. and Maryland, are comprised of over 100 series with motion picture films, drawings of incarceration centers, photos, maps, correspondence, yearbooks, rosters, etc.

Series 51, the “Internal Security Case Reports” from 1942 to 1946, comprises narrative reports prepared by camp investigators, police officers, and directors of internal security, relating cases of alleged “disorderly conduct, rioting, seditious behavior,” etc. at each of the 10 camps, with detailed information on the names and addresses in the camps of the persons involved, the time and place where the alleged incident occurred, an account of what happened, and a statement of action taken by the investigating officer.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Collecting Data</td>
<td>Using Computational Models to Understand a Concept</td>
<td>Preparing Problems for Computational Solutions</td>
<td>Investigating a Complex System as a Whole</td>
</tr>
<tr>
<td>B. Creating Data</td>
<td>Using Computational Models to Find and Test Solutions</td>
<td>Programming</td>
<td>Understanding the Relationships within a System</td>
</tr>
<tr>
<td>C. Manipulating Data</td>
<td>Assessing Computational Models</td>
<td>Choosing Effective Computational Tools</td>
<td>Thinking in Levels</td>
</tr>
<tr>
<td>D. Analyzing Data</td>
<td>Designing Computational Models</td>
<td>Assessing Different Approaches/Solutions to a Problem</td>
<td>Communicating Information about a System</td>
</tr>
<tr>
<td>E. Visualizing Data</td>
<td>Constructing Computational Models</td>
<td>Developing Modular Computational Solutions</td>
<td>Defining Systems and Managing Complexity</td>
</tr>
<tr>
<td>F.</td>
<td></td>
<td>Creating Computational Abstractions</td>
<td></td>
</tr>
<tr>
<td>G.</td>
<td></td>
<td>Troubleshooting and Debugging</td>
<td></td>
</tr>
</tbody>
</table>
A. Creating Data

“The increasingly computational nature of working with data in” archival science “underscores the importance of developing computational thinking practices in the classroom.” “Part of the challenge is teaching students that answers are drawn from the data available.” “In many cases” archivists “use computational tools to generate data… at scales that would otherwise be impossible.”

<table>
<thead>
<tr>
<th>Japanese Name</th>
<th>Last Name</th>
<th>First Name</th>
<th>Anglo Name</th>
<th>Incident Date</th>
<th>Year</th>
<th>Age</th>
<th>Residence ID</th>
<th>Family Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Abe</td>
<td>Kiyotake</td>
<td>Joe</td>
<td>11-4-43</td>
<td>1943</td>
<td></td>
<td>2904-D</td>
<td>40562</td>
</tr>
</tbody>
</table>

Diagram:

1. Index Card Image
2. OCR
3. Index Card Text
4. NER
5. Metadata Extracted from Card

Image:

- Index Card showing handwritten notes and a card with printed information.
<table>
<thead>
<tr>
<th>Last Name</th>
<th>First Name</th>
<th>Birth Year</th>
<th>Original State</th>
<th>Gender</th>
<th>Birth Place</th>
<th>Family No</th>
<th>Individual No</th>
<th>File Number</th>
<th>Assembly Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABE</td>
<td>FRANK</td>
<td>1910</td>
<td>CA</td>
<td>M</td>
<td>CA</td>
<td>24067</td>
<td>24067A</td>
<td>208156</td>
<td>None</td>
</tr>
<tr>
<td>ABE</td>
<td>FRANK</td>
<td>1940</td>
<td>CA</td>
<td>M</td>
<td>CA</td>
<td>24067</td>
<td>24067</td>
<td>201888</td>
<td>None</td>
</tr>
<tr>
<td>ABE</td>
<td>FRANK</td>
<td>1905</td>
<td>CA</td>
<td>M</td>
<td>Honolulu county</td>
<td>8605</td>
<td>08605A</td>
<td>950783</td>
<td>Fresno</td>
</tr>
<tr>
<td>ABE</td>
<td>FRANK</td>
<td>1913</td>
<td>CA</td>
<td>M</td>
<td>Oregon</td>
<td>18050</td>
<td>18050B</td>
<td>805536</td>
<td>Santa Anita</td>
</tr>
<tr>
<td>ABE</td>
<td>FRANK</td>
<td>1881</td>
<td>CA</td>
<td>M</td>
<td>Sakhalin</td>
<td>34424</td>
<td>34424C</td>
<td>207865</td>
<td>None</td>
</tr>
</tbody>
</table>

WRA Form 26 register
"Japanese-American Internee Data File"
NARA AAD

Box 8 -- #269

3-23-45 A-1067 INFRAC. PROJ. REG.

ABE, Frank Tomo
1514-A

The above was put in project jail for military marching, blowing of bugles, display of Japanese emblems. Occurred in the colony.

<table>
<thead>
<tr>
<th>#</th>
<th>Last</th>
<th>First</th>
<th>Family No</th>
<th>Sex</th>
<th>Birth</th>
<th>Citizenship</th>
<th>Alien #</th>
<th>Entry</th>
<th>Entry Date</th>
<th>Pre-evacuation Addr</th>
<th>Type of Final Departure</th>
<th>Date of Final Departure</th>
<th>Destination of Final Departure</th>
</tr>
</thead>
<tbody>
<tr>
<td>96</td>
<td>ABE</td>
<td>Tomo</td>
<td>18050</td>
<td>M</td>
<td>2-1-13</td>
<td>R</td>
<td>-</td>
<td>Granada, CO</td>
<td>9-18-43</td>
<td>Los Angeles, CA</td>
<td>Dept. of Justice Internment</td>
<td>6-24-45</td>
<td>Santa Fe, NM</td>
</tr>
</tbody>
</table>

Final Accountability Rosters (FAR)
B. Manipulating Data

“Computational tools make it possible to efficiently and reliably manipulate large and complex” archival holdings. “Data manipulation includes sorting, filtering, cleaning, normalizing, and joining disparate datasets.”

<table>
<thead>
<tr>
<th>Japanese Name</th>
<th>Last Name</th>
<th>First Name</th>
<th>Other Name</th>
<th>Incident Date</th>
<th>Year</th>
<th>Age</th>
<th>Residence ID</th>
<th>Family Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Abe</td>
<td>James</td>
<td></td>
<td>10/5/1942</td>
<td>1942</td>
<td>20</td>
<td>2803-A</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>Abe</td>
<td>Makoto</td>
<td></td>
<td>10/7/1942</td>
<td>1943</td>
<td>43</td>
<td>1206-A</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>Abe</td>
<td>Sakichi</td>
<td></td>
<td>11/2/1942</td>
<td>1942</td>
<td>62</td>
<td>5315-B</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>Abe</td>
<td>Shigeki</td>
<td></td>
<td>5/1/1943</td>
<td>1943</td>
<td>43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>Abe</td>
<td>Kiyotake</td>
<td>Joe</td>
<td>11/4/43</td>
<td>1943</td>
<td></td>
<td>2904-D</td>
<td>40562</td>
</tr>
<tr>
<td>N</td>
<td>Jensen</td>
<td>Lloyd</td>
<td>H.</td>
<td>11/16/43</td>
<td>1943</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
C. Analyzing Data

“There are many strategies that can be employed when analyzing data for use in” an archival context, “including looking for patterns or anomalies, defining rules to categorize data, and identifying trends and correlations.”

- We used NER software to extract metadata from the incident cards. This was done with the open source GATE. This is based on pattern matching through recognition rules. The matching rules are often refined through iterative tuning.

- For example, a rule for recognizing a person’s name would be based on a lastname, followed by a comma, followed by a Japanese firstname, followed by an Anglo first name in parentheses. As we process additional cards we would note that there are other styles of names, so the pattern would be generalized account for stylistic variations. If the pattern is made to be robust enough it will eventually work on all of the instances of names.

- GATE, General Architecture for Text Processing, [https://gate.ac.uk/](https://gate.ac.uk/)
D. Visualizing Data

“Communicating results is an essential component of” understanding archival data “and computational tools can greatly facilitate that process. Tools include both conventional visualizations such as graphs and charts, as well as dynamic, interactive displays.”

Box 8  
WRA Form 26  
FAR Tule LAke
E. Designing Computational Models

“The ability to create, refine, and use models of phenomena is a central practice.” “Models can include flowcharts and diagrams.” “Part of taking advantage of computational power... is designing new models that can be run on a computational device.” “There are many reasons that might motivate designing a computational model, including wanting to better understand a phenomenon under investigation, to test out a hypothesis.” “Students... will be able to define the components of the model, describe how they interact, decide what data will be produced by the model.”
An important practice is the ability to create new or extend existing computational models. This requires being able to encode the model features in a way that a computer can interpret.

FOR EACH of the 10,000 cards:
    IF cardname is Japanese:
        MATCH cardname in the FAR registry
        IF MATCH is true:
            COMPUTE date difference
            IF difference > 18:
                RELEASE card
            ELSE:
                DO NOT RELEASE card: possible PII
        ELSE:
            MATCH cardname in the Form 26 registry
            IF MATCH is true:
                COMPUTE date difference
                IF difference > 18:
                    RELEASE card
                ELSE:
                    DO NOT RELEASE card: possible PII
            ELSE:
                PII Status NOT DETERMINED
    ELSE:
        RELEASE card: WRA Staff – Not Internee
“Enabling students to explore” archival problems “using computational problem solving practices such as programming, algorithm development, and creating computational abstractions.” “The ability to encode instructions in such a way that a computer can execute them is a powerful skill for investigating” archival problems. Programs include ten-line Python scripts.

```python
for i in range( 0, 10000 ):
    if Internee( cardName ):
        print_cardInfo( i )
    if card_Name in FAR_list:
        FAR_lookup(i)
        PII_DateCheck(i)
        if age(i) > 18:
            print( 'RELEASABLE' )
        else:
            print( 'NO RELEASABLE' )
    elif card_Name in FORM26_list:
        FORM26_lookup(i)
        PII_DateCheck(i)
        if age(i) > 18:
            print( 'RELEASABLE' )
        else:
            print( 'NO RELEASABLE' )
    else:
        print( 'NOT DETERMINED' )
    else:
        print( 'RELEASABLE' )
```
H. Developing Modular Computational Solutions

“When working toward a specific archival outcome, there are often a number of steps or components involved in the process; these steps, in turn, can be broken down in a variety of ways that impact their ability to be easily reused, repurposed, and debugged. Developing computational solutions in a modular, reusable way has many implications. By developing modular solutions, it is easier to incrementally construct solutions, test components independently, and increase the likelihood that components will be useful for future problems.”

- We make use of abstraction and functional programming through the use of modular components such as:
  - `PII_DateCheck()`,
  - `FORM26_lookup()`, and
  - `FAR_lookup()`.

- This allows for reusable chunks of code that can be tested locally. The larger program is the composition of these modules, which makes it both more readable and maintainable.
I. Creating Computational Abstractions

"The ability to create and use abstractions is used constantly across Archival Science undertakings, be it creating computational abstractions when writing a program, generating visualizations of data to communicate an idea or finding, defining the scope or scale of a problem, or creating models to further explore or understand a given phenomenon. Creating computational abstractions is essential for solving multiple problems that have structural similarity but differ in surface detail."

○ The class of Japanese Person Names having attributes surname, given name, and anglo name is a data abstraction.

○ The instances of this concept have different forms in the text of index cards
  ○ Surname alone
  ○ Given name alone
  ○ Anglo name alone
  ○ Surname given name
  ○ Surname, given name
  ○ And others
J. Troubleshooting and Debugging

“Troubleshooting broadly refers to the process of figuring out why something is not working or behaving as expected. There are a number of strategies one can employ while troubleshooting a problem, including clearly identifying the issue, systematically testing the system to isolate the source of the error, and reproducing the problem so that potential solutions can be tested reliably.”

- To facilitate group debugging, we use an interactive server-based shared version of Jupyter Notebook.
  - “The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.”

- Project Jupyter, see: http://jupyter.org/
The InterPARES Preservation Model

- The InterPARES Preservation Model provides a generic preservation strategy (or framework) for preserving authentic electronic records.
- Within that framework, a variety of preservation strategies can be developed.
- The Preservation Model provides a framework for implementing procedures that satisfy the Authenticity Task Force's (ATF's) Baseline Requirements for Supporting the Production of Authentic Copies of Electronic Records.
Investigating a Complex System as a Whole

“Investigating a complex system as a whole relies on the ability to define and measure inputs and outputs of the system. … Being able to black box the details of the underlying systematic interactions and focus on the system as a whole makes it possible to understand the characteristics of the system in aggregate…Computational tools such as models and simulations are especially useful in such investigations,”
Understanding the Relationships within a System

“Thus, it is important to be able to identify the different elements of a system and articulate the nature of their interactions. Computational tools are useful for conducting such inquiry as they can provide learners with controls for isolating different elements, investigating their behaviors, and exploring how they interact with other components of the system.”

To refine and validate the Preservation Model, a walkthrough using case study data was performed.

The method used in the walkthrough is to iteratively step through each of the lowest-level activities in the Preservation Model:

1. Reviewing the activity definition and the input, output and control definitions.
2. Identifying data elements of labels on input and output arrows.
3. Defining the transformation of inputs to outputs.
4. Determining values of the data elements that are related to the specific body of records in the case study.
5. Recording the results and any problems or issues that arise and suggesting possible solutions of the model was conducted using information from a case study.
Thinking in Levels

“Systems can be understood and analyzed from different perspectives, ranging from a micro-level view that considers the smallest elements of the system to a macro-level view that considers the system as a whole. Thinking about a system from the standpoint of its individual actors and components can lead to insights about how micro-level behaviors lead to emergent macro-level patterns.”
Communicating Information about a System

“Communicating information about a system often involves developing effective and accessible visualizations and infographics that highlight the most important aspects of what has been learned about the system in such a way that it can be understood by someone who does not know all the underlying details.”
Defining Systems and Managing Complexity

“The decision of where to set the boundaries of the system is critical for any investigation that follows as it determines what questions you can answer as well as the size and complexity of the system.”
The Authenticity Task Force of the InterPARES Project investigated the essential attributes of authentic digital records and the technological and procedural factors impacting their authenticity.

They formulated the conceptual requirements for presuming the authenticity of the digital records maintained by a records creator, which they terms Benchmark Requirements.

The benchmark requirements draw specifically on the notion of a trusted record-keeping system.

They formulated a method for assessing a presumption of authenticity of records based on the degrees of belief that the Benchmark Requirements are supported by evidence.
Benchmark Requirements

To support a presumption of authenticity, the preserver must obtain evidence that:

<table>
<thead>
<tr>
<th>REQUIREMENT A.1: Expression of Record Attributes and Linkage to Record</th>
</tr>
</thead>
<tbody>
<tr>
<td>the value of the following attributes are explicitly expressed and inextricably linked to every record. These attributes can be distinguished into categories, the first concerning the identity of records, and the second concerning the integrity of records.</td>
</tr>
<tr>
<td><strong>A.1.a</strong> Identity of the record:</td>
</tr>
<tr>
<td><strong>A.1.a.i</strong> Names of the persons concurring in the formation of the record, that is:</td>
</tr>
<tr>
<td>• name of author</td>
</tr>
<tr>
<td>• name of writer (if different from the author)</td>
</tr>
<tr>
<td>• name of originator (if different from name of author or writer)</td>
</tr>
<tr>
<td>• name of addressee</td>
</tr>
<tr>
<td><strong>A.1.a.ii</strong> Name of action or matter</td>
</tr>
<tr>
<td><strong>A.1.a.iii</strong> Date(s) of creation and transmission, that is:</td>
</tr>
<tr>
<td>• chronological date</td>
</tr>
<tr>
<td>• received date</td>
</tr>
<tr>
<td>• archival date</td>
</tr>
<tr>
<td>• transmission date(s)</td>
</tr>
<tr>
<td><strong>A.1.a.iv</strong> Expression of archival bond (e.g., classification code, file identifier)</td>
</tr>
<tr>
<td><strong>A.1.a.v</strong> Indication of attachments</td>
</tr>
<tr>
<td><strong>A.1.b</strong> Integrity of the record:</td>
</tr>
<tr>
<td><strong>A.1.b.i</strong> Name of handling office</td>
</tr>
<tr>
<td><strong>A.1.b.ii</strong> Name of office of primary responsibility (if different from handling office)</td>
</tr>
<tr>
<td><strong>A.1.b.iii</strong> Indication of types of annotations added to the record</td>
</tr>
<tr>
<td><strong>A.1.b.iv</strong> Indication of technical modifications:</td>
</tr>
</tbody>
</table>
## Benchmark Requirements

<table>
<thead>
<tr>
<th>REQUIREMENT A.2: Access Privileges</th>
<th>the creator has defined and effectively implemented access privileges concerning the creation, modification, annotation, relocation, and destruction of records;</th>
</tr>
</thead>
<tbody>
<tr>
<td>REQUIREMENT A.3: Protective Procedures: Loss and Corruption of Records</td>
<td>the creator has established and effectively implemented procedures to prevent, discover, and correct loss or corruption of records;</td>
</tr>
<tr>
<td>REQUIREMENT A.4: Protective Procedures: Media and Technology</td>
<td>the creator has established and effectively implemented procedures to guarantee the continuing identity and integrity of records against media deterioration and across technological change;</td>
</tr>
<tr>
<td>REQUIREMENT A.5: Establishment of Documentary Forms</td>
<td>the creator has established the documentary forms of records associated with each procedure either according to the requirements of the juridical system or those of the creator.</td>
</tr>
<tr>
<td>REQUIREMENT A.6: Authentication of Records</td>
<td>if authentication is required by the juridical system or the needs of the organization, the creator has established specific rules regarding which records must be authenticated, by whom, and the means of authentication;</td>
</tr>
<tr>
<td>REQUIREMENT A.7: Identification of Authoritative Record</td>
<td>if multiple copies of the same record exist, the creator has established procedures that identify which record is authoritative;</td>
</tr>
<tr>
<td>REQUIREMENT A.8: Removal and Transfer of Relevant Documentation</td>
<td>if there is a transition of records from active status to semi-active and inactive status, which involves the removal of records from the electronic system, the creator has established and effectively implemented procedures determining what documentation has to be removed and transferred to the preserver along with the records.</td>
</tr>
</tbody>
</table>
Using Computational Models to Understand a Concept

“Computational models that demonstrate specific ideas or phenomena can serve as powerful learning tools. Students can use computational models to deepen their understanding of degrees of belief.

- Although the method of assessment is expressed in simple terms, there are substantial pitfalls inherent in subjective probability assessment due to psychological biases and common misunderstandings of probabilistic reasoning.

- Furthermore, the conditional dependencies between requirements and between the evidence needed to conclude that a requirement is met can be quite complex.

- It was suggested that the Bayesian approach to reasoning with degrees of belief combined with Bayesian Belief Networks could overcome some of the common psychological biases and fallacies in reasoning due to misunderstanding of probability.

- Hence, the original research question might be reformulated as:
  - Can the method for assessment of authenticity based on the Benchmark Requirements be more precisely specified and tested using Bayesian Probability and Bayesian Belief Networks so that a preserver could be confident that he could apply the method and Requirements and be confident in the result?
Using Computational Models to Find and Test Solutions

“Computational models can also be used to test hypotheses and discover solutions to problems. They make it possible to test many different solutions quickly, easily, and inexpensively before committing to a specific approach. This is especially helpful for phenomena whose outcomes depend on multi-dimensional “parameter spaces.”
Assessing Computational Models

“A key practice in using a computational model effectively is to understand how the model relates to the phenomenon being represented. This understanding is guided by a variety of questions including: Which aspects of the phenomenon have been faithfully modeled and which aspects have been simplified or ignored? What assumptions have the creators of the model made about the world and how do those assumptions affect its behavior? What layers of abstraction have been built into the model itself and how do these abstractions shape the fidelity of the model?”

- The Baseline requirements and their conditional dependencies were represented as a Bayesian Belief Network using the GeNIe graphical user interface.
- A number of conditional dependencies in the Benchmark Requirements were exposed.
- Other factors influencing the Benchmark requirements such as user authentication and denial of service to non-users were identified.
- An approach to testing the BBN model was described using data from a recordkeeping system for scientific data records.
Automated Categorization of Email for Records Retention

○ Every USA federal agency is required to manage its records according to a record retention schedule that requires the classification of email as records.

○ The increasing volume of email makes it impractical to manually categorize email.

○ The objective of this project was to investigate whether machine learning techniques could achieve a high degree of accuracy in automatically categorizing email.

○ An experiment was conducted in which support vector machine classifiers were trained to classify six categories of actual GTRI email.

○ The six classifiers were then evaluated in classifying email not in the training set.

○ The result was 99% accuracy in classifying the additional emails.
Preparing Problems for Computational Solution

“While some problems naturally lend themselves to computational solutions, more often, problems must be reframed so that existing computational tools can be utilized. Strategies for doing this include decomposing problems into subproblems, reframing new problems into known problems for which computational tools already exist, and simplifying complex problems so the mapping of problem features onto computational solutions is more accessible.”

- Traditionally, determining the filing category of paper records involved human decision.
- The increasing volume of email makes it impractical to manually categorize email.
- The problem is reframed as developing a computational tool, supervised machine learning, to solve this problem.
Collecting Data

“Data are collected through observation and measurement. Computational tools play a key role in gathering and recording a variety of data across many different archival endeavors. Computational tools can be useful in different phases of data collection, including the design of the collection protocol, recording, and storage.”

<table>
<thead>
<tr>
<th>Filing Category</th>
<th>Retention Category</th>
<th>Samples</th>
<th>Training Samples</th>
<th>Test Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty Library Advisory Board Email</td>
<td>(A4) Advisory Board Records</td>
<td>101</td>
<td>68</td>
<td>33</td>
</tr>
<tr>
<td>GTRI Director Administrative Email</td>
<td>(A13) Correspondence, Administrative</td>
<td>134</td>
<td>90</td>
<td>44</td>
</tr>
<tr>
<td>ACL ListServ Email</td>
<td>(A15) Correspondence, Transitory</td>
<td>100</td>
<td>65</td>
<td>35</td>
</tr>
<tr>
<td>Employee Benefits Email</td>
<td>(A16) Correspondence, General</td>
<td>67</td>
<td>48</td>
<td>19</td>
</tr>
<tr>
<td>ITTL Brown Bag Email</td>
<td>(A38) Special Event Records</td>
<td>77</td>
<td>45</td>
<td>32</td>
</tr>
<tr>
<td>Help Desk Computer System Maintenance Email</td>
<td>(D1) Computer System Maintenance Records</td>
<td>98</td>
<td>63</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>577</td>
<td>379</td>
<td>198</td>
</tr>
</tbody>
</table>
Choosing Effective Computational Tools

“A single task can often be solved a number of different ways using a variety of different computational tools. In such cases, there is often a single tool, or at least a small subset of tools, for the job. Being able to identify the strengths and weaknesses of various possible tools for the problem at hand can be the most important decision in a project. Choosing an effective computational tool includes considering the functionality it provides, its scope and customizability, the type of data the tools expects and can produce, as well as questions that extend beyond the software itself, such as, whether or not there is an active user community that could assist with difficulties you might encounter. “

○ Text categorization is the problem of assigning a selected document to one or more categories.

○ There are two primary approaches to automated text categorization—rule-based and statistics-based (machine learning).

○ Rule-based text processing tools used for classification tasks promise very high recall/precision or accuracy values. The documents are classified by hand-made rules. The whole process brings about high costs in analyzing and modeling the application domain.

○ Supervised machine learning trains classifiers on a set of documents that have been labeled with the correct class. Given a sufficient sample for each category, machine learning models generally cost less to create than rule-based systems. They are also easier to scale up to large volumes of email. The supervised machine learning methods that have been used for text categorization include: Maximum Entropy classification; Naïve Bayes; and Support Vector Machines (SVMs).

○ Supervised Machine Learning promises low costs both in analyzing and modeling the application at the expense of a lower accuracy. It is independent of domain specific knowledge.
Assessing Different Approaches/Solutions to a Problem

“When there are multiple approaches to solving a problem or multiple solutions to choose from, it is important to be able to assess the options and make an informed decision about which route to follow. Even if two different approaches produce the same, correct result, there are other dimensions that should be considered when choosing a solution or approach, such as cost, time, durability, extendibility, reusability, and flexibility.”

○ Support Vector machines are one of the most effective supervised machine learning methods for text categorization.

○ SVMs:
  ○ scale up to high dimensionalities (number of features),
  ○ work well without term (feature) selection, and
  ○ generalize well to other data, i.e., are robust to over-fitting
Conclusions and Next Steps

- Computational Thinking is an important facet of Archival Science Research.
- It is essential that computational thinking be included as part of MLIS Archival Studies education.
- Can we develop computational thinking enhanced lesson plans for archival topics that could be used by iSchool faculty to introduce computational thinking into their courses?
- Can we build an international networked community of iSchool faculty and Library and Archives practitioners to engender these capabilities?