Using NLP to Support Dynamic Arrangement, Description, and Discovery of Born Digital Collections:

The ArchExtract Experiment

Mary Elings
Principal Archivist for Digital Collections
Bancroft Library, UC Berkeley
NLP trains computers to process and understand human language

Includes:

- Topic Modelling
- Keyword Extraction
- Named Entity Extraction
The Theory

– Can text analysis/NLP aid archivists in arranging and describing large text-based archival collections?
– Can these tools be automated for use by non-technical users?
Environmental Scan

• NLP projects or tools focused on archival text collections
  – 2013-2014: TOME (Interactive TOpic Model and MEtadata Visualization), Georgia Tech
  – 2013-2015 ePADD, Stanford University Libraries
Mallet
machine learning for language toolkit

C:\>ed mallet
C:\mallet>bin\mallet
Mallet 2.0 commands:
    import-dir           load the contents of a directory into mallet instances (one
                        per file)
    import-file          load a single file into mallet instances (one per line)
    import-svmlight      load a single SVMLight format data file into mallet instance
    train-classifier     train a classifier from Mallet data files
    train-topics         train a topic model from Mallet data files
    infer-topics         use a trained topic model to infer topics for new documents
    estimate-topics      estimate the probability of new documents given a trained model
    hlda                 train a topic model using Hierarchical LDA
    prune                remove features based on frequency or information gain
    split                divide data into testing, training, and validation portions
    Include --help with any option for more information

C:\mallet>bin\mallet import-dir --input appleton --output appleton.mallet --remove-stopwords
Labels =
        appleton
C:\mallet>
The equipartition theorem is a formula from statistical mechanics that relates the temperature of a system with its average energies. The original idea of equipartition was that, in thermal equilibrium, energy is shared equally among its various forms. For example, the average kinetic energy in the translational motion of a molecule should equal the average kinetic energy in its rotational motion. Like the...
ArchExtract’s Goals:

- web-based
- require no knowledge of the command-line or programming
- create a standard semi-automated workflow for archivists
- provide varying levels of access to users in different roles
ArchExtract Implementation

A ruby on rails web-application that:

- packages and manages a number of topic modeling, named entity and keyword algorithms
- provides interface to browse and explore the text collection
Process Flow for a Collection

Import

Preprocess

Extract Named Entities

Run topic modeling

Extract Keywords

Review results

Publish for researchers to view
Import Collection

Can import pdfs, word docs, and excel files
Dec 5, 1874
Delta Chambers
Liverpool
My dear Cousin Muir!
I duly received your favour of the 14th ult yesterday along with your first book which I handed over to my wife last evening for which we both thank you very much. She picked out about the squirrels the very first time read it out aloud to some friends who happened to be present. Well, now, about another matter. We have been in regular receipt of the San Francisco Bulletin tho.
United States Senate,
WASHINGTON, D.C.
April 9, 1917.

My dear Boys:

This is the first of the letters I am now writing in our new office. I enclose you, merely as a momento of the occasion, copy of the address of the President last night and copy of the amended resolution declaring war. The first indication of action that I have observed was this morning when the amended resolution was reported by the Foreign Affairs Committee. Mr. La Follette asked that consideration, under the rules, be held over until tomorrow. Mr. Martin, the Democratic leader, wished to indulge either in a lecture or denunciation of La Follette in a sentence or two that he uttered to the effect that we must not appeal to us with our controversial natures and our acerbity for scraps which had the applause of the Democratic side of the Senate and of the galleries. Tomorrow, there ought to be some brief pyrotechnics and then the declaration of war.
### Ladies' Relief Society Records

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>25</td>
<td>1</td>
<td>1</td>
<td>31</td>
<td>5</td>
<td>30</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>25</td>
<td>2</td>
<td>2</td>
<td>29</td>
<td>12</td>
<td>30</td>
<td>12</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>30</td>
<td>3</td>
<td>3</td>
<td>35</td>
<td>10</td>
<td>30</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>April</td>
<td>28</td>
<td>2</td>
<td>1</td>
<td>33</td>
<td>16</td>
<td>30</td>
<td>16</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>May</td>
<td>34</td>
<td>1</td>
<td>1</td>
<td>31</td>
<td>15</td>
<td>30</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>June</td>
<td>34</td>
<td>1</td>
<td>1</td>
<td>31</td>
<td>15</td>
<td>30</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>July</td>
<td>34</td>
<td>1</td>
<td>1</td>
<td>31</td>
<td>15</td>
<td>30</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>August</td>
<td>34</td>
<td>1</td>
<td>1</td>
<td>31</td>
<td>15</td>
<td>30</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>September</td>
<td>29</td>
<td>2</td>
<td>1</td>
<td>33</td>
<td>16</td>
<td>30</td>
<td>16</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>October</td>
<td>25</td>
<td>1</td>
<td>1</td>
<td>25</td>
<td>15</td>
<td>20</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>November</td>
<td>25</td>
<td>1</td>
<td>1</td>
<td>25</td>
<td>15</td>
<td>20</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>December</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>25</td>
<td>15</td>
<td>20</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Preprocess

Reduce overall size of the collection vocabulary

- stop words
- words that only appear once
- part of speech tagging
- tf-idf filtering
- stem words
Run a PreProcessing Job on a Collection

- Select Collection
  - Collection: [Input Field]

- Stop Word Filtering
  - Stop Words: Remove common words like the, a, it
    - Yes
    - No

- Remove Rare Words: Remove words that only appear once in the document
  - Yes
  - No

- Import a Custom Stopword: Remove words that appear in the uploaded list
  - Browse: [File Upload]

- Tagging: Mark Names/Phrases

- Tagging: Include OHU: the part of speech words selected
  - Nouns
  - Priors
  - Adjectives
  - Proper Names

- Remove Named Entities When Tagging: Remove words that are identified as named entities and data
  - Yes
  - No

- Stemming

- Stem: Turn each word into its stem
  - Yes
  - No

- TF-ICF Filtering

  - Note: This option cannot be combined with stemming or tagging
  - TF-ICF: Filter words based on their importance for analysis. Words are ranked in descending order of frequency, frequency of occurrence and frequency of occurrence outside documents

- Filter out words that are below the mean 1st degree for the entire collection
Run Topic Modeling

- implements mallet to produce topics
- can produce 1-100 topics
- each topic shows the documents that are most closely associated with the topic
- no single method – users can use different pre-processes and number of topics to find best results
John Muir Collection: Plain Text With 50 Topics

You can delete topics by clicking on the x

× Topic 0: day sun light days wind air began snow dark strange hours winds cold round storm clear fell till storms
× Topic 1: day morning evening saturday leave kt sunday tomorrow train night afternoon letterhead monday hotel meet sellers week yesterday friday
× Topic 2: river mountains grand canon region mountain sierra mt forest forests trip south california king north kings canyon summer made
× Topic 3: yosemite valley park mr made road men sheep present influence people state irish noble part management strong report general
× Topic 4: foot tree trees hundred high pine ground ft big side forest grove twenty miles thousand long north fine large
× Topic 5: god love world heart life man give words word lord joy spirit bless pray live blessed things mother heaven
× Topic 6: san california francisco st sir car give pacific date remain address july kindly favor state library respectfully reply september
Topic 2: river mountains grand canon region mountain sierra mt forest forests trip south california king north kings canyon summer made

Top documents associated with this topic:

- john_muir_papers_pdf_kt5s204057 → 0.46718
- john_muir_papers_pdf_kt467mfr → 0.25144
- john_muir_papers_pdf_kt9d09s2db → 0.22778
- john_muir_papers_pdf_kt4g5036eh → 0.21984
- john_muir_papers_pdf_kt0v20r906 → 0.19007
- john_muir_papers_pdf_kt5199r8e → 0.19209
- john_muir_papers_pdf_kt0s03120 → 0.18056
- john_muir_papers_pdf_kt9o9r801 → 0.17666
- john_muir_papers_pdf_kt5m9fn1mr → 0.1749
- john_muir_papers_pdf_kt8d5nz2c7 → 0.17204
- john_muir_papers_pdf_kt2t1nf10k → 0.16725
Dear Mr. Muir,

I am sending you herewith for your information copy all introduced by me to at add the King-Kern addition to the sequoia National Park in California as proposed by the Sierra Club of California. I am likewise enclosing a copy of a letter received by me today from the Secretary of the Interior with reference to the bill. I assure you that I have been very glad to be of service to those interested in this matter.

Yours truly,

[Signature]

Mr. John Muir,
325 West Adams Street,
Los Angeles, California. Enc.

04696 DEPARTMENT OF THE INTERIOR
Input Outcomes

Topic Results (Best to Worst)
• Born digital text
• Hand transcribed text
• Uncorrected OCR from printed text

The larger the collection, the better the topics
Named Entities
uses the Stanford named entity recognizer

Keywords
implements algorithms that use the python natural language toolkit library (nltk)

Group similar words together using fuzzy matching
## Named Entities for the John Muir Collection

<table>
<thead>
<tr>
<th>Top 50 People</th>
<th>Top 50 Organizations</th>
<th>Top 50 Places</th>
<th>Top 50 Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NAME</strong></td>
<td><strong>COUNT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kern</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hooper</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carroll</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>David Muir</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Almond</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sam</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thompson</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>John Your</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Douglas</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McKinley</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrenhard</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J. Muir</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adams</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Le Conte</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>James D. Butler</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perry</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agassiz</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>John Nolan</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wanda Muir</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J. L. Hudson</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anna W. Chaney</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baker</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>William F. Bade</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albright</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edwin H. Moulton</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bryan</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maggie Lunnem</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kent</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. Hort Mannum</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bartlett</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arthur</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>King</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Davis</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harper</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W. W. Hannan</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nawahiem</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duncan</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Julia M. Moores</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clara Barnas</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fay</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chas</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>William</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lacey</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mrs Hooker</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marjorie</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annie K. Bidwell</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McGlone</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taylor</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thorne</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goucher</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
LeConte

Similar terms:
 Co, Conte, LeConte, J. N. LECONTE, Joseph LeConte, JOSEPH N. LECONTE, Leconte, Joseph N. LeConte, John LeConte, J. N. LeConte, Leconte, J. LeConte, Joe LeConte, Conte, N. J. LeConte, O. BRYANT PROF. JOSEPH N. LECONTE, S. F. Prof. J. N. LECONTE, Helen M. LaConte, Joe LaConte, EC

Appears 12 times in the following documents:

john_muir_papers.pdf h0779e4q1 → 1
john_muir_papers.pdf k06008f2t → 1
john_muir_papers.pdf k9008m3nd3nr → 1
john_muir_papers.pdf k178b60e5ns → 1
john_muir_papers.pdf k1267mn86n → 1
john_muir_papers.pdf k4599nn7ht → 1
john_muir_papers.pdf k108v18s1c2 → 1
john_muir_papers.pdf k15649sh0mm → 1
john_muir_papers.pdf k1067md837 → 1
john_muir_papers.pdf k11149r36r → 1
john_muir_papers.pdf k1681m0w1 → 1
john_muir_papers.pdf k4220r8p2 → 1

Back
• after reviewing output, can publish topics and named entities
• give researchers and archivist different roles in the web application
• assigned role determines the view and functionality users can access
Conclusions

- Can text analysis/NLP aid archivists in describing large text-based archival collections?
  
  *Yes, with caveats*

- Can these tools be automated for use by non-technical users?
  
  *Yes, but more testing needs to be done*
Thank you!

Check out the project on github at: http://github.com/j9recurses/archextract

Email: melings@berkeley.edu