



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

Natural Language Processing

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
 - 1998: Greenberg, Jane, UNC-CH. “The Applicability of Natural Language Processing (NLP) to Archival Properties and Objectives.” *American Archivist*
 - 2001-2012: Underwood, William. Georgia Tech
 - 2013: Thomas Padilla, UIUC (MSU): “Topic Modelling Archival Materials.” *Practical E-Records*
 - 2013-2014: TOME (Interactive TOPic Model and METadata Visualization), Georgia Tech
 - 2013-2014: Ed Summer, MITH: *Fondz*
 - 2013-2015 ePADD, Stanford University Libraries



```
Command Prompt

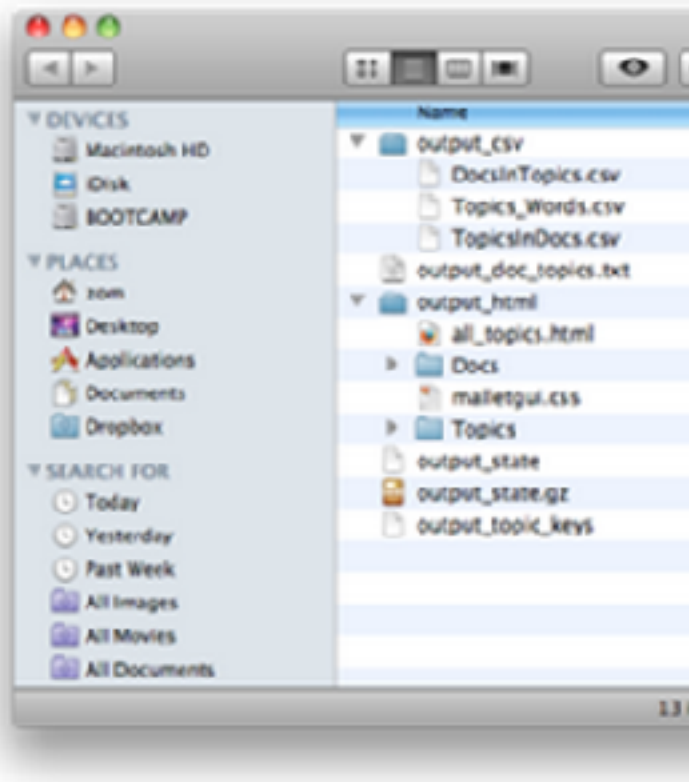
C:\>cd mallet

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
e (one per line)
  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 mo
del
  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 --remo
ve-stopwords
Labels =
  appleton
C:\mallet>
```



Topic Modelling Tool



The screenshot shows a web browser window displaying the output of the Topic Modelling Tool. The browser address bar shows the URL: http://www.khourya.com/technology/info/output_html/Docs/Doc1.html. The page title is 'Doc1.html'. The main content area displays the following information:

DOC: equipartition_theorem.txt

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

Top-4 topics in this doc (% words in doc assigned to this topic)

- (71%) system average equipartition theorem energy kinetic drama richard considered effects stars classical heat motion equilibrium thermal energies premier boyfriend ...
- (7%) sunderland echo paper edward uranian related thought survived daily storey areas east newspaper evening april states temperature alvida romance ...
- (5%) tasmanian london confederates century relative devil species thylacinus greek headed december thomas forced australia return don appeared publications composed ...
- (5%) gunnild norway law life sullivan gilbert thespis death top discovered actors gods creating johnston leading batsman shiloh rulers rest ...

[Index]

ArchExtract's Goals:

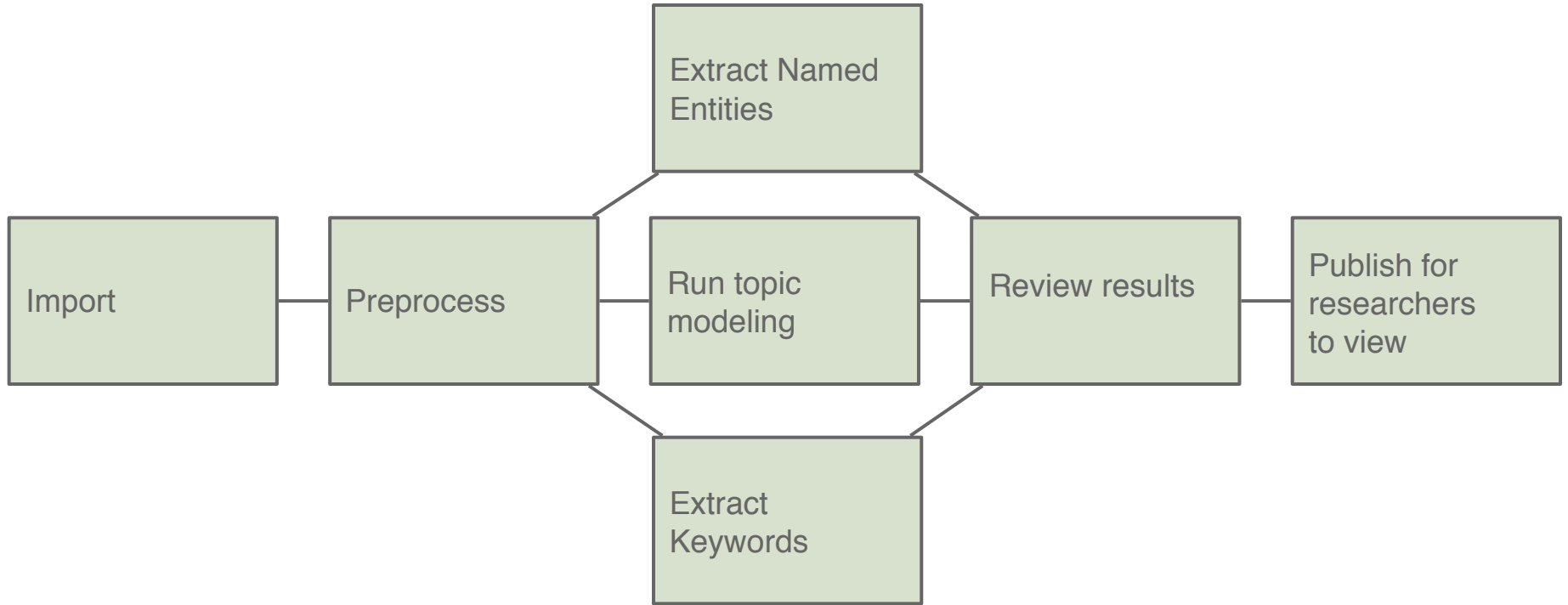
- web-based
- require no knowledge of the command-line or programming
- create a standard semi-automated workflow for archivists
- provide varying levels of access to users in different roles

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 Collection

Can import pdfs, word docs, and excel files

243
4 the 12 pins of Connemara
they had strange adventures
sometimes reduced to star-
vation fare & perishing cold
but never a days illness.
Promising prentices for
John Muir the mountaineer
truly! But my dear fellow
you must really give it
up at your time of life &
be content to leave the
Lochs & Blessed wilderness
& its further exploration
to some body else who
can afford to be killed without
having to leave a widow & her
bonnie bairnies behind them.
With all our united kindest
love to thee & thine I am ever
your affectionate cou-
sine
John Muir Esq
Martinez
California } Jas M. Hay

247
Dec. 4. 1894
Delta Chambers
Liverpool.
My dear Cousin Muir.
I duly received your
favour of the 14th ult yester-
day 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.
We are now about another

Hand transcribed text

John Muir Correspondence

Transcription:

4 the 12 pins of Connemara they had strange adventures sometimes reduced to starvation fare perishing cold but never a days illness. Promising prentices for John Muir the mountaineer truly But my dear fellow you must really give it up at your time of life be content to leave the illegible Blessed wilderness its further exploration to some body else who can afford to be killed without having to leave a widow her bonnie bairnies behind them. With all our united kindest love to thee thine I am ever Your affectionate cou-
sine
Jas M. Hay John Muir Esq Martinez California 1 Dec. 4. 1894 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 the 01884

United States Senate,
WASHINGTON, D. C.

April 3, 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 memento 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 Commission. 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

21 Civitob Slixler, role,
WASHINGTON, D. C.
April 3, 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 memento 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 Commission, 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 (sentences that would not have appealed to us with our controversial nature and our alacrity 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.

Uncorrected OCR from
(clean-ish) printed text

*Hiram Johnson
Correspondence*

Born digital text

Ladies' Relief Society Records

Statistik T. Snow Home
Monthly Census
Year 2001

Month	Census End 1	March 1	Transfer	Census End of Month	Census Quarts	Census 100%	Gain	Loss	Overseer (no. students)	Capacity
January	21	1	1	21	25%	25%	0			23
February	21	2	2	27	27%	27%	0	2		23
March	23	3	3	35	25%	25%	7			23
April	23	3	1	23	25%	25%	3			23
May	18	3	2	18	25%	25%	9	1		23
June	31	3	2	35	25%	25%	6	1		23
July	24	3	4	27	25%	25%	10	1		23
August	27	3	3	25	25%	25%	9			23
September	23	2	2	23	25%	25%	9			23
October	24	1	1	25	25%	25%	7	1		23
November	23	1	1	25	25%	25%	9	1		23
December										

The screenshot shows a digital spreadsheet application with a grid layout. The columns contain numerical data, and the rows are organized by month. Several rows are highlighted in blue, likely representing specific categories or totals. The interface includes a menu bar at the top and a status bar at the bottom.

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 a Collection

* Collections

Select a Collection

Stop Word Filtering

Removes Stop Words: Removes common words like *the, a, it*

- Yes
 No

Removes Rare Words: Removes words that only appear once in the document

- Yes
 No

Applied a Custom Stoplist: Removes words that appear in the uploaded list

[Browse...](#) No file selected.

Tagging Part of Speech Tagging

Tagging: Include ONLY the part of speech words selected

- Nouns
 Verbs
 Adjectives
 Proper Names

Removes Named Entities When Tagging: Removes words that are identified as named entities and dates

- Yes
 No

Stemming

Stem: Turn each word into its stem

- Yes
 No

Tf-IDF Filtering

Note: This option cannot be combined with stemming or tagging

Tf-IDF: Filter words based on their importance in the document. Common words will be the minimum, words that frequently do occur together and have independently the word appears in other documents

Filter out words that are below the **min-tf-idf** score 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

search topics

← Previous **1** 2 3 4 5 Next →
[Go Back to Topic Model Home](#)

You can delete topics by clicking on the x

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

document → Model topic share score

john_muir_papers_pdf_kt9s204057 → 0.46718

john_muir_papers_pdf_kt2n35r7kf → 0.33049

john_muir_papers_pdf_kt467nf0fr → 0.26144

john_muir_papers_pdf_kt8c6036ck → 0.23585

john_muir_papers_pdf_kt9x09s3zb → 0.22778

john_muir_papers_pdf_kt6d5nf3hd → 0.22517

john_muir_papers_pdf_kt4g5034sh → 0.21984

john_muir_papers_pdf_kt829037xn → 0.21171

john_muir_papers_pdf_kt0r23r606 → 0.19807

john_muir_papers_pdf_kt5q2nf370 → 0.19526

john_muir_papers_pdf_kt5199r8s1 → 0.19209

john_muir_papers_pdf_kt6g5034v5 → 0.18519

john_muir_papers_pdf_kt05803120 → 0.18056

john_muir_papers_pdf_kt4d5nf1qp → 0.17683

john_muir_papers_pdf_kt5z09r801 → 0.17566

john_muir_papers_pdf_kt6h4nf2z2 → 0.17507

john_muir_papers_pdf_kt5m3nf1mr → 0.17449

john_muir_papers_pdf_kt8f59s0h1 → 0.17293

john_muir_papers_pdf_kt6d5nf2c7 → 0.17204

john_muir_papers_pdf_kt2t1nf10t → 0.16735



Current Topic: river mountains grand canon region mountain sierra mt forest forests trip south california king north kings canyon summer made

Original Document Contents for john_muir_papers_pdf_k19s204057

k19s204057

Illegible

MR. TALIAFERRO, MR. SMITH, MD., MR. SIMMONS, MR. HUGHES, MR. JOHNSTON,
SOL. N. SHERIDAN, CLERK, CARL V. KING, ASST. CLERK.

United States Senate
COMMITTEE ON INTER-OCEANIC CANALS.

March 1, 1911.

My dear Mr. Muir:

I am sending you herewith for your information copy of bill introduced by me to set aside the Kings-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 to-day 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,
Illegible

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

04696 DEPARTMENT OF THE INTERIOR

Plain Text with 50 Topics

Top 50 Topics Associated with this Document:

CURRENT TOPIC:

river mountains grand canon
region mountain sierra mt
forest forests trip south
california king north kings
canyon summer made

→ 0.46716

ADDITIONAL TOPICS:

• matter interior states public
forest proposed washington
government united city report
reserve secretary lands land
department purpose reservation
national → 0.15315

• feet tree trees hundred high
pine ground ft big side forest
grove twenty miles thousand long
north fine large → 0.11869

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

Extract Named Entities and Keywords

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 Entites for the John Muir Collection

Top 50 People

Top 50 Organizations

Top 50 Places

Top 50 Dates

NAME → COUNT

X Kerr → 14	X Wanda Muir → 12	X Duncan → 12
X Hooper → 13	X J. L. Hudson → 12	X Julia M. Moores → 12
X Camoll → 13	X Anna W. Cheney → 12	X Clara Barnus → 12
X David Muir → 13	X Baker → 12	X Fay → 12
X Alvord → 13	X William F. Bade → 12	X Chas → 11
X Sam → 13	X Alchison → 12	X Williams → 11
X Thompson → 13	X Edwin F. Moulton → 12	X Lacey → 11
X John Year → 13	X Bryson → 12	X Mrs Hooker → 11
X Douglas → 13	X Maggie Lunam → 12	X Marjorie → 11
X Mc Kinley → 13	X Kent → 12	X Annie K. Bidwell → 11
X Washburn → 13	X G. Hart Merriam → 12	X Mc Clane → 11
X J. Muir → 13	X Bartlett → 12	X Taylor → 11
X Adams → 13	X Arthur → 12	X Thoreau → 11
X La Conte → 12	X King → 12	X Goucher → 11
X James D. Butler → 12	X Davis → 12	
X Perry → 12	X Harper → 12	
X Agassiz → 12	X W. W. Hannan → 12	
X John Nolan → 12	X Newsham → 12	



LeConte

THE JOHN MUIR COLLECTION

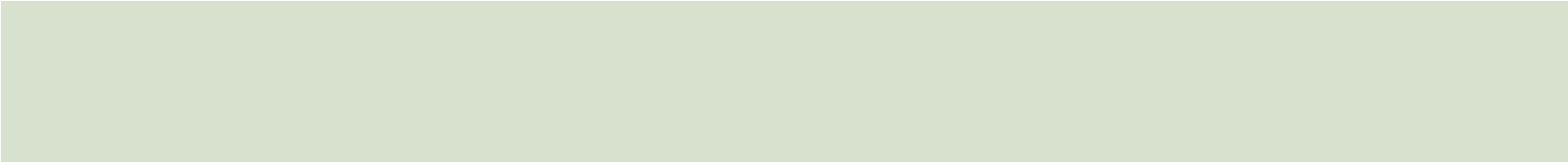
Similar terms:

Co, Conte, LeConte, J. N. LÉCONTE, Joseph LeConte, JOSEPH N. LÉCONTE, Lecante, Joseph M. LeConte, John LeConte, J. N. LeConte, Leonte, J. LeConte, Joe LeConte, Conte, N, Joe LeConte, G. BRYANT PROF. JOSEPH N. LÉCONTE, S. F. Prof. J. M. LÉCONTE, Helen M. LeConte, Joe. LeConte, EC

Appears 12 times in the following documents:

john_muir_papers_pdf_k0779e4q → 1
john_muir_papers_pdf_k090002ot → 1
john_muir_papers_pdf_k00m3nd5r → 1
john_muir_papers_pdf_k08b69s0sz → 1
john_muir_papers_pdf_k0267nd8nn → 1
john_muir_papers_pdf_k0038nd7ht → 1
john_muir_papers_pdf_k00v18s1c2 → 1
john_muir_papers_pdf_k09j49s0nn → 1
john_muir_papers_pdf_k0067nd837 → 1
john_muir_papers_pdf_k11j49c0r → 1
john_muir_papers_pdf_k0611n0w1 → 1
john_muir_papers_pdf_k4z05r8p2 → 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