Background

- IMLS Research Project
  - Improving Search and Discovery Using Topic Modeling
  - Yale (lead), UMICH, UC Irvine

- Apply topic modeling to three classes of digital library resources: full-text books, images, and tagged objects
- Build prototypes of user interfaces that make use of topics
- Test the prototypes to assess the value of topic modeling for users
Collections and challenges

• Digitized books
• Images
• Scientific literature
• Web 2.0 content
• … and more
Collections and challenges

• Digitized books

• Images

• Scientific literature

• Web 2.0 content

• … and more

Currently Digitized

• 6,182,629 total volumes
  3,621,100 book titles
  146,505 serial titles
  2,163,920,150 pages
  230 terabytes
  73 miles
  5,023 tons
Collections and challenges

- Digitized books
- Images
- Scientific literature
- Web 2.0 content
- … and more

- Catalog Search
  - Subj: “American Colonial History” 20,000 results
- Full-Text Search
  - “American Colonial History” 1,000,000 results
- Limitation
  - Users don’t have mental model
  - Users don’t trust metadata
Collections and challenges

- Digitized books
- Images
- Scientific literature
- Web 2.0 content
- … and more
Collections and challenges

- Digitized books
- Images
- Scientific literature
- Web 2.0 content
- ... and more
Collections and challenges

• Digitized books
• Images
• Scientific literature
• Web 2.0 content
• … and more

• 1000 new articles daily
• Indexed using MeSH
What is Topic Modeling?

- Topic Modeling (aka Latent Dirichlet Allocation)
- Updated version of Latent Semantic Analysis
- State-of-the-art model for collections of text documents
- Works great on large collections of well written content
Topic model learns topics from co-occurring words

Think of topic modeling as automatic assignment of subject headings … that you learn

Collection of NSF Awards

**Topic Model:**
- Use words from title & abstract
- Learn 400 topics

“Topic Tags”

<table>
<thead>
<tr>
<th>t49</th>
<th>t18</th>
<th>t114</th>
<th>t305</th>
</tr>
</thead>
</table>

... topics...

t13. particles particle colloidal granular material ...
t14. ocean marine scientist cosee oceanography ...
t15. atmospheric chemistry ozone air organic ...
...
• Gain an understanding of topic modeling
• Understand where topic modeling might add value
• Learn how to integrate topics into the search back-end
• See how to use topic modeling user interfaces
Collections

- UMich History of ART
  - example
  - 100,000s images
- Books from Internet Archive
  - link
  - 100,000s books
Workflow

Collection

Topic model

topics

Topics in each doc

Solr

Solr

Solr

Solr
Topic Modeling

• Download MALLET
  – http://mallet.cs.umass.edu/

• Follow instructions to build

• Topic model collection
  1. import-dir
  2. train topics
Topic Modeling

1. **mallet import-dir**
   - --input /data/mycoll
   - --output mycoll.mallet.in
   - --keep-sequence
   - --remove-stopwords

2. **mallet train-topics**
   - --input mycoll.mallet.in
   - --num-topics 20
   - --output-topic-keys topics.txt
   - --output-doc-topics topicsindocs.txt
1. `mallet import-dir`
   - `--input /data/mycoll`
   - `--output mycoll.mallet.in`
   - `--keep-sequence`
   - `--remove-stopwords`
   - `--extra-stopwords mystopwords.txt`

2. `mallet train-topics`
   - `--input mycoll.mallet.in`
   - `--num-topics 20`
   - `--output-topic-keys topics.txt`
   - `--output-doc-topics topicsindocs.txt`
• Review topics.txt

• Review topicsindocs.txt
• Download Solr

• Follow instructions to build

• Ingest collection into solr
  1. Convert to “xml”
  2. Use post.sh
Solr

- Follow example in exampledocs directory
- Review solr.xml
- Post.sh solr.xml
- Check in solr admin panel
- Run searches in solr admin panel
Solr

- Solr is the popular, blazing fast open source enterprise search platform from the Apache Lucene project. Its major features include powerful full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling. Solr is highly scalable, providing distributed search and index replication, and it powers the search and navigation features of many of the world's largest internet sites.
Researchers in the C2 (Creative Consilience of Computing and the Arts) Program at Yale are investigating *Sketching and Alternative Design Techniques* in architecture:

- **Computer graphics plays a major role in the architecture profession.** For example, modeling and rendering systems have proven to be invaluable aids in the visualization process, allowing designers to walk through their designs with photorealistic imagery. However, computer graphics techniques are typically employed at the conclusion of the design process. In fact, most of the artistic and intellectual challenges of an architectural design have already been resolved by the designer sits down in front of a computer.

- [http://topics.catalog.hathitrust.org](http://topics.catalog.hathitrust.org)
- [http://quod.lib.umich.edu/cgi/i/image/image-idx?c=hart4topics](http://quod.lib.umich.edu/cgi/i/image/image-idx?c=hart4topics)
SPARE SLIDES
A closer look at one automatically learned topic

*topic-6*: conflict violence war international military domestic political government terrorism national security civil …

- What is this topic about? Is it a meaningful topic?

- [How] Do we present this to users? … What is a good label for this topic?
Overarching Questions

Q1: Are individual topics meaningful and usable?

Q2: Are assignments of topics to documents meaningful and usable?

Q3: Do topics facilitate better or more efficient document search, navigation, browsing?
## Experimental Setup

<table>
<thead>
<tr>
<th>Collection</th>
<th>Sources</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>Internet Archive</td>
<td>12,000 books</td>
</tr>
<tr>
<td></td>
<td>Hathi Trust</td>
<td>28,000 books</td>
</tr>
<tr>
<td>News Articles</td>
<td>LDC Gigaword (NY Times articles)</td>
<td>55,000 articles</td>
</tr>
<tr>
<td>Grant Abstracts</td>
<td>National Institutes of Health</td>
<td>60,000 grants</td>
</tr>
<tr>
<td>Image Metadata</td>
<td>Yale Library</td>
<td>100,000s</td>
</tr>
<tr>
<td></td>
<td>UMich Library</td>
<td>100,000s</td>
</tr>
</tbody>
</table>
Experimental Setup

- Topic modeled each document collection (using different topic resolutions). Selected a total of 600 topics for manual coherence scoring.

- Have N = 9-15 annotators score each of the 600 topics on a 3-point scale where 3 = “useful” (coherent) and 1 = “useless” (less coherent), based on the top-10 topic words.
  - also asked annotators to identify “best” topic word … and
  - suggest a short label.
Coherent (unanimous score=3)

Books

silk lace embroidery tapestry gold embroidery
trout fish fly fishing water angler stream

Less coherent (unanimous score=1)

News

space earth moon science scientist light nasa

Health drug patient medical doctor hospital care

Metadata

japan scroll kamakura ink hanging oyobe hogge silk

persian iran manuscript folio firdawsi century

abstraction sculpture united female english tabbaa

oil john canvas william james england henry robert

Incoherent topics are not errors ... they are also statistical patterns of word usage seen in the data
Automatic Scoring of Topics?

- Coherence of topic depends on relatedness of all 10-choose-2 pairs of top-10 topic words

- Idea: Use external data to evaluate word pair relatedness (e.g. Wikipedia)
Relatedness of word pairs

Topic: music dance band rock opera …

Pointwise Mutual Information (measure of dependence)

Count co-occurrence in a sliding window

Dance music works often bear the name of the corresponding dance, e.g. waltzes, the tango, the bolero, the can-can, minuets, salsa, various kinds of jigs and the breakdown. Other dance forms include contradance, the merengue (Dominican Republic), and the cha-cha-cha. Often it is difficult to know whether

the name of the music came first or the name of the dance.

#(dance,music) = 1
Relatedness of word pairs

**Topic:** music dance band rock opera …

Pointwise Mutual Information (measure of dependence)

\[ PMI(w_1, w_2) = \log \frac{Pr(w_1, w_2)}{Pr(w_1)Pr(w_2)} \]

\[ PMI-\text{Score}(w) = \sum_{ij} PMI(w_i, w_j), ij \in 1\ldots10, i < j \]
Relatedness of word pairs

Topic: music dance band rock opera ...

PMI-Score = 4.5 + 4.2 + ... + 1.4
PMI-score achieves human-level performance

BOOKS (280 topics, corr=0.78)
PMI-score achieves human-level performance

NEWS (117 topics, corr=0.77)
<table>
<thead>
<tr>
<th>Collection</th>
<th>Human-human correlation</th>
<th>PMI-Human correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>0.76</td>
<td>0.78</td>
</tr>
<tr>
<td>News Articles</td>
<td>0.73</td>
<td>0.77</td>
</tr>
<tr>
<td>Grant Abstracts</td>
<td>0.48</td>
<td>0.63</td>
</tr>
<tr>
<td>Image Metadata</td>
<td>0.51</td>
<td>0.53</td>
</tr>
</tbody>
</table>
Outliers

• PMI-score over-predicts coherence
  – thou thy thee hast hath thine mine heart god heaven
  – viii vii xii xiii xiv xvi xviii xix xvii main
  – century fifteenth thirteenth fourteenth twelfth sixteenth middle.
  – want look going deal try bad tell sure feel remember

• PMI-score under-predicts coherence
  – public government america policy nation political issues …
  – british britain england country united national foreign nation …
  – account cost item profit balance statement sale credit loss …
<table>
<thead>
<tr>
<th>Topic</th>
<th>Suggested Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>trout fish fly</td>
<td>fly fishing</td>
</tr>
<tr>
<td>fishing</td>
<td></td>
</tr>
<tr>
<td>water angler stream rod flies</td>
<td></td>
</tr>
<tr>
<td>space</td>
<td>space exploration</td>
</tr>
<tr>
<td>earth moon science scientist light nasa</td>
<td></td>
</tr>
<tr>
<td>race car</td>
<td>nascar racing</td>
</tr>
<tr>
<td>nascar</td>
<td></td>
</tr>
<tr>
<td>driver racing</td>
<td></td>
</tr>
</tbody>
</table>
Best topic word task

Topic

tROUT fish fly fishing water angler stream rod flies ...

space earth moon science scientist light nasa ...

race car nascar driver racing ...

Features
• PMI( word1, word2 )
• Prob( word | * ) … word is evoked by other words … e.g. space
• Prob( * | word ) … evokes other words … e.g. nascar

SVM-rank using above features beats baseline of first topic word (Lau, Newman, Karimi, Baldwin COLING 2010)
<table>
<thead>
<tr>
<th>Topic</th>
<th>Wiki Article Titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>trout fish fly fishing water angler stream rod flies</td>
<td>fly fishing fishing angling trout</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>space earth moon science scientist light nasa</td>
<td>space exploration space science</td>
</tr>
<tr>
<td></td>
<td>space colonization nasa missions</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
Large-Scale User Studies

- Developed prototype user interfaces for Image Collections and Book Collections that use topics

- Large scale user studies at Yale and UMich underway

- Assessing qualitative and quantitative value of topics
Prototype with topic facets

Catalog Search

american colonial history

Narrow Search

Viewability
- Limited (search-only) (523)
- Full view (31)

Topic
- architecture (84)
- cultural identity (66)
- slavery (51)
- furniture (44)
- politics (30)

Subject
- Architecture United States History 20th century (29)
- United States (26)
- Art American 20th century Exhibitions
What else we can do with topics?

Visualization of search results (each dot is a search result)
Topic trends throughout books

[SENTIMENT] felt comfort feeling feel spirit mind heart point moment ill letter beyond mother state never event evil fear impossible hope time idea left situation poor distress possible hour end loss relief dearest suffering concern dreadful misery unhappy emotion …
Conclusion

- Topic models seem to be useful in digital libraries for creating additional metadata …

- … but learned topics can vary in usefulness and coherence

- We developed model to automatically evaluate topic coherence that matches human judgments

- This is a step in integrating topic modeling into digital libraries
Thank You
Goal

• Improve Search and Discovery
  – Improve user experience

• How
  – Improve search results
  – Improve ranking of search results
  – Improve display of search results
Collections and problems

- The internet (web pages)
- Scientific literature
- News articles
- Digitized books
- Images

- Uses PageRank to rank search results
- We’ve learned to use it
- We’re used to it
- Limitation: One size fits all
Collections and challenges

• The internet (web pages)
• Scientific literature
• News articles
• Digitized books
• Images
Collections and problems: FIXME: DELETE

- The internet (web pages)
- Scientific literature
- News articles
- Digitized books
- Images

- How to get cross-cutting topics?
- Limitation: cross-cutting topics?
Evaluation

- Statistical
- User
  - How to eval
  - TREC-style
Selected high-scoring topics (unanimous score=3):
[News] space earth moon science scientist light nasa mission planet mars ...
[News] health disease aids virus vaccine infection hiv cases infected asthma ...
[Books] steam engine valve cylinder pressure piston boiler air pump pipe ...
[Books] furniture chair table cabinet wood leg mahogany piece oak louis ...

Selected low-scoring topics (unanimous score=1):
[News] king bond berry bill ray rate james treas byrd key ...
[News] dog moment hand face love self eye turn young character ...
[Books] soon short longer carried rest turned raised lled turn allowed ...
[Books] act sense adv person ppr plant sax genus applied dis ...
Topic modeling can relate and unify documents from different collections

Topic Model:
• Automatically learns set of topics for any collection of documents
• Labels each document with a few topics
• Provides a way to relate documents from different collections
• Reveals flow of ideas
• Is mature technology and considered state-of-art
• Is basis for this tool!
• Measure Spearman rank correlation between different methods and the average of the user ratings (reversing the sign for distance-based methods)

• Upper bound = average inter-annotator correlation, as measured by leave-one-out cross validation between the annotators