Evaluating Topic Models for Digital Libraries

Kat Hagedorn, University of Michigan
David Newman, University of California, Irvine
Youn Noh, Yale University

Yale University Library
4 October 2011
Talk Outline

- Introduction (Youn)
- Topic Evaluation (Dave)
- Topic Assignment (Youn)
- Topics in Applications (Kat)
- Questions and Discussion (All)
INTRODUCTION
Project Scope, Purpose, and Participants

- **Scope**: Research and development project funded by a National Leadership Grant from the Institute of Museum and Library Services
- **Purpose**: Investigate applications of topic modeling for library and museum collections
- **Participants**:
  - Yale University (lead)
  - University of California, Irvine
  - University of Michigan
Project Goals

- Propose topic modeling as one solution to organizing the massive number of digital objects produced by large scale digitization projects
- Measure the extent to which topic modeling improves search and discovery
- Disseminate these results and findings
- Show how topic modeling can be deployed and applied by multiple institutions and constituencies
- Produce open source software and courseware for others to use
Ranganathan’s Five Laws of Library Science

1. Books are for use.
2. Every reader his [or her] book.
4. Save the time of the User.
5. The library is a growing organism.

Maybe topic modeling could help!

Image: Some rights reserved by Sarah_G
How can you get involved?

- Ask questions or get in touch after the talk
- Attend the Teaching with Technology Tuesdays session today at 11:00 in Bass L01
- Contact us
  - Youn Noh ([youn.noh@yale.edu](mailto:youn.noh@yale.edu))
  - David Newman ([newman@uci.edu](mailto:newman@uci.edu))
  - Kat Hagedorn ([khage@umich.edu](mailto:khage@umich.edu))
Acknowledgments

Adrian Knight * Arun Balagopalan
Bill Dueber * Carolyn Caizzi * Elizabeth Boyce
Emmanuelle Delmas-Glass * Ernie Marinko
Francesca Slade * Holly Rushmeier * Huang Huang
Joan Swanekamp * John Weise * Julie Dorsey
Justo Arosemena * Karen Kupiec * Karen Rizvi
Katie Bauer * Ken Hamma * Marlon Forrester
Meg Bellinger * Megha Okhai * Michael Kargela
Mike Friscia * Nick Garver * Pam Franks
Patrick Paczkowski * Rebekah Irwin * Robert Carlucci
Robert Nelson * Roger Espinosa * Scott Matheson
Scott Wilcox * Sherlock Campbell * Suzanne Chapman
Evaluating Topic Models
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

$q = \text{“madonna and child”}$
Collections and challenges

- Digitized books
- Images
- Scientific literature
- Web 2.0 content
- ... and more
- 1000 new articles daily
- Indexing 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

Collection of NSF Awards

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

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

```
t49
t18
t114
t305
```

```
t13. particles particle colloidal granular material …
t14. ocean marine scientist cosee oceanography …
t15. atmospheric chemistry ozone air organic …
```

"Topic Tags"

Evaluating Topic Models
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 Hathi Trust</td>
<td>12,000 books</td>
</tr>
<tr>
<td></td>
<td></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 UMich Library</td>
<td>100,000s</td>
</tr>
<tr>
<td></td>
<td></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 for each topic.
Example Coherent and Incoherent Topics

Coherent  
(unanimous score=3)

Books
silk lace embroidery tapestry gold embroidered ...
trout fish fly fishing water angler stream rod flies ...

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 ...

Less coherent  
(unanimous score=1)
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)

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

\[ PMI - Score(w) = \sum_{ij} PMI(w_i, w_j), ij \in 1 \ldots 10, i < j \]
Relatedness of Word Pairs

Topic: music dance band rock opera …

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

NEWS (117 topics, corr=0.77)
## Experiments

<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 …
### Best topic word and suggested Label

<table>
<thead>
<tr>
<th>Topic</th>
<th>Suggested Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>trout fish fly <em>fishing</em> water angler stream rod flies ...</td>
<td>fly fishing</td>
</tr>
<tr>
<td><em>space</em> earth moon science scientist light nasa ...</td>
<td>space exploration</td>
</tr>
<tr>
<td>race car <em>nascar</em> driver racing ...</td>
<td>nascar racing</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**

- $\text{PMI}(\text{word1}, \text{word2})$
- $\text{Prob}(\text{word} | * )$ … word is evoked by other words … e.g. space
- $\text{Prob}( * | \text{word} )$ … 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>space earth moon science scientist light nasa ...</td>
<td>space exploration space space science space colonization nasa missions ...</td>
</tr>
</tbody>
</table>
Large-Scale User Studies

- Developed prototype user interfaces for Images Collections and Books Collections that use topics

- Large scale user studies at Yale and UMich underway

- Qualitative and quantitative assessment value of topics
Conclusion

• Topic models seem to be useful in DLs for creating additional metadata

• But learned topics can vary in quality and coherence

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

• This is a useful step in integrating topic modeling into digital libraries
TOPIC ASSIGNMENT
Overview (1)

• We frequently have to categorize things in our everyday lives (email, web pages, documents, photos)

• But this task is often time-consuming and error-prone and, for library collections, requires specialized training
• Recall that topic modeling simultaneously learns the topics in a collection and assigns topics to documents

• Topic modeling is similar to open card sorting

• In an open card sort, documents are sorted into conceptual categories that haven’t been pre-specified

• For topic modeling:
  – The # of topics (or categories) is the only input
  – Documents (or cards) may belong to more than one topic (or category)
Overview (3)
Latent topics are revealed in the sorting process

In a separate step, categories are labeled based on the scope and content of cards in the category

In a similar fashion, topics are labeled based on the scope and meaning of the words in the topic
Pros and Cons of Topic Modeling

**Pros**
- Automatically generates topics and assignments
- Takes relatively little time
- Topics are consistently and comprehensively assigned
- Easy to update topics and assignments as collection grows and its scope evolves

**Cons**
- Topics are lists of words that need to be labeled
- Domain knowledge is often required to interpret and label topics
- Statistically valid topics are not necessarily meaningful to users
• **Purpose of study**: Explore potential for using learned topics as subjects

• **Research question**: How well do learned topics perform, as compared to Library of Congress Subject Headings (LCSH), in clustering like documents?

• **Practical application**: Many digital library applications use categories or related items
Scope of Study

- **Participants**: 88 students from the University of Michigan and 62 students from Yale University
- **Test Collection**: 27,710 digitized books in the subject areas of art and architecture from the University of Michigan Libraries
Study Design

- 21 matched topic-LCSH pairs
- 5 books assigned each topic or LCSH
A Sample Pair

muscle bone lower head upper tendon joint front arm finger
form surface outer rib spine ...

Anatomy, Artistic

The Human figure, its beauties and defects
Ernst Wilhelm von Brücke

Figure construction
Alon Bement

Anatomical diagrams for the use of art students
James M. Dunlop
drawer leg chair chest top mahogany front rail seat walnut carved furniture cupboard cabinet pine ...
A Sample Question

Do these books belong together?

Strongly disagree 1 2 3 4 5 Strongly agree
Test Format

- Testing was conducted online
- Each of the 3 versions of the test had 14 items (7 topic-LCSH pairs presented in randomized order)
- In addition to the cover, participants could look inside the book, using a Google Preview button
Results

- We received 47-50 responses for each item.
- We applied the Mann Whitney U test to the mean scores of the 21 topic-LCSH pairs:
  - 11 (p<.05) received higher topic scores.
  - 4 (p<.05) received higher LCSH scores.
  - 6 (p>.05) received similar scores.
- Overall, topics outperformed LCSH.
Observations

• Some topics brought together words used in different contexts
  – A sample topic: *air* heating *pipe* heat *gas* hot *supply* temperature *tank* boiler *steam equipment* electric *pressure* fuel ...
  – Assigned to books on garages, public baths, theaters, and office buildings

• Topic assignments were generally more specific than LCSH assignments, so the books in the topic sets tended to be more cohesive
Topics Used in Applications

Kat Hagedorn
University of Michigan Libraries
Background

• Using a topic modeling algorithm to find relevant materials in a large corpus of textual items is not new.

• However, to date there has been little investigation into its usefulness to end-users.
Investigation: Two Methods

- Unmoderated study
  - no one to help participants!
  - complicated setup & process
- Moderated study
  - face-to-face discussions
  - experts w. their own searches

Do you want $15 for 15 minutes of your time?

Please go up the stairs to the Library's 2nd floor reference area and look for the computer with the orange arrow on top.

You will receive a $15 Amazon gift certificate for your participation.

This study is part of a grant between the University of Michigan, Yale University and the University of California Irvine. It is funded by the Institute of Museum and Library Services.
HathiTrust: special interface

• Used a subset of the HathiTrust Digital Library to test both methods.
  – 27,710 digitized volumes
  – corresponding to 18,881 records
  – from the "N" (fine arts) Library of Congress Subject Heading (LCSH) classification
  – primarily in the English language
  – linked to both in-copyright and full-view-access volumes
HathiTrust: special interface
Unmoderated: testing script

• Tasks and questions designed to utilize unmoderated testing environment by allowing participant to freely find results that suited the task.

• At the same time, required them to describe the results and records they found by answering follow-up questions.
Unmoderated: testing script

• 1st task: perform a general search for urban planning, as if you were part of an introductory architecture class
• 2nd task: do a different search in the same subject area and this time use the Narrow Search column
• 3rd task: do a different search and specifically focus on the topic facet
• Satisfaction vectors (from the end-of-task surveys):
  – results they found
  – book they chose
  – facets column
  – topics in facets column
Unmoderated: setup

- Completely hands-off testing.
- The sign was a huge help!

Paper printout was vital to guide participants...
Unmoderated: setup

- Because Morae was not actually an out-of-the-box tool, needed support services to build layers around it.
Unmoderated: underway

Participants could limit by topic in the left-hand (Narrow Search) column

This participant narrowed by the topic “architecture”
Unmoderated: underway

Participants stepped through the tasks in Morae
Unmoderated: underway

At the end of each task (3 total), there was a survey.

The survey was the only method we could use to find out if participants were satisfied with their results.
Unmoderated results

• There were a lot of results! (306 participants)
• Some results were contradictory...

• Overall... usage of topic facets stayed high and constant throughout the study. Between 73-83% of participants used the topic facets.
## Contradictory results

Satisfaction rating went down after participants were asked to use the topic facets specifically... but stayed high when participants were asked to look at all facets.

<table>
<thead>
<tr>
<th>Task 2 questions (when asked to look at all facets)</th>
<th>Task 2 satisfaction range</th>
<th>Task 3 questions (when asked to look specifically at topic facets)</th>
<th>Task 3 satisfaction range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Were you satisfied with the book you chose?</td>
<td>77% satisfied</td>
<td>Were you satisfied with the book you chose?</td>
<td>50% satisfied</td>
</tr>
<tr>
<td>Were you satisfied with the results you found?</td>
<td>71% satisfied</td>
<td>Were you satisfied with the results you found?</td>
<td>35% satisfied</td>
</tr>
<tr>
<td>Did you find the &quot;Narrow Search&quot; column on the left useful in refining your results?</td>
<td>78% satisfied</td>
<td>Did you find the &quot;Topics&quot; listed in the &quot;Narrow Search&quot; column on the left useful in refining your results?</td>
<td>44% satisfied</td>
</tr>
</tbody>
</table>
Contradictory results

- Participants who used more topic facets had higher satisfaction scores, when they were not specifically asked to use them.

<table>
<thead>
<tr>
<th>Topic Facets Selected</th>
<th>Average Satisfaction Score Task 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.410714</td>
</tr>
<tr>
<td>1</td>
<td>1.704545</td>
</tr>
<tr>
<td>2</td>
<td>1.800000</td>
</tr>
</tbody>
</table>

Seems that use of, and satisfaction with using, the topics is tied to the use of other facets in the facet column.

1=high satisfaction, 5=low satisfaction
Moderated: testing script

• Keen to try a different methodology: replicate results or discover variations on them.
• Since we had been unable to discuss anything with participants during the original study, face-to-face testing was an obvious next step.
• Also, unmoderated study due to its test design had focused in large part on the use of the facets themselves and to a small extent on the content of the created topics.
Moderated: testing script

- **1\textsuperscript{st} task**: compare two searches, one using topic facets to limit the search
- **2\textsuperscript{nd} task**: comment on quality and ranking of search performed using your own search terms; compare LCSH & topic facets
- **3\textsuperscript{rd} task**: look at a record from this search; compare LCSH & topic facets; comment on appropriateness of topics
Moderated results

Topics were helpful although LCSH was considered superior to topics. However, large majority of participants thought topics assigned to records were appropriate.

<table>
<thead>
<tr>
<th>Task</th>
<th>Quality of results</th>
<th>Superiority / helpfulness of topics</th>
<th>Appropriateness of topics in records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1: review 2 sets of results (our search)</td>
<td>8 of 8 (100%) = good</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Task 2: review topics and LCSH in results (their search)</td>
<td>5 of 8 (62.5%) = good 3 of 8 (37.5%) = poor</td>
<td>4 of 8 (50%) = topics superior 3 of 8 (37.5%) = LCSH superior 1 of 8 (12.5%) = neither</td>
<td>n/a</td>
</tr>
<tr>
<td>Task 3a: choose and review topic within results (their search)</td>
<td>6 of 8 (75%) = good 2 of 8 (25%) = poor</td>
<td>5 of 8 (62.5%) = topics helped 3 of 8 (37.5%) = topics didn’t help</td>
<td>n/a</td>
</tr>
<tr>
<td>Task 3b: review topics and LCSH in record (their search)</td>
<td>n/a</td>
<td>6 of 8 (75%) = LCSH superior 1 of 8 (12.5%) = topics superior 1 of 8 (12.5%) = neither</td>
<td>7 of 8 (87.5%) = topics appropriate 1 of 8 (12.5%) = topics not appropriate</td>
</tr>
</tbody>
</table>
Moderated results

- Most interesting result... when participants looked at LCSH and topics together, 50% found topics superior and 37.5% found LCSH superior = about equal.

- Participants say:
  - "They [topics and LCSH] seem to relate well to each other."
  - "It [usefulness throughout the study] did switch between topics and subject [LCSH]. So in that sense it's great to have both."

Seems like conjunction of the two types of subject-based limiters is helpful to participants hoping to further refine a search result.
More support for topics...

• Asked participants if they thought the use of topics gave accurate results (books chosen matched topics chosen). 87.5% said yes.

• Asked participants if they would use topic facets again in this interface. 75% said yes.

• Also... some results support conclusion that using more than one topic at a time to narrow or expand a search can be more effective than using just one topic on its own. (And this effectiveness decreases when using more than three topics at a time.)
Conclusions

• Unmoderated study told us: use of, and satisfaction with using, the topics is tied to the use of other facets in the facet column.

• Moderated study told us: conjunction of the two types of subject-based limiters is helpful to participants hoping to further refine a search result.

• We can conclude that participants view topics and LCSH as apples and apples, not apples and oranges. They want to use them together to refine their results.
Further Work

• What results would we get using all of the records in the Summon service from Serials Solutions?

• What results would we get if we topic modeled all of the HathiTrust volumes (currently over 9.5 million)?

• How would the results differ if we used a collection of Law or Astronomy texts?

• How would users evaluate topics against a different type of classification such as the Art & Architecture Thesaurus (AAT)?
Further information

• Kat works at University of Michigan Libraries
  <http://www.lib.umich.edu/lit/dlps>

• Current (September/October) issue of D-Lib Magazine has full article on this work
  <http://dx.doi.org/10.1045/september2011-hagedorn>