Learning Topics and Related Passages in Books

David Newman  
Computer Science  
Univ. California, Irvine  
neuman@uci.edu

Youn Noh  
Office of Digital Assets and Infrastructure  
Yale University  
youn.noh@yale.edu

Kat Hagedorn  
University Libraries  
University of Michigan  
khage@umich.edu

Arun Balagopalan  
Computer Science  
Univ. California, Irvine  
abalagop@ics.uci.edu

ABSTRACT
The number of books available online is increasing, but user interfaces may not be taking full advantage of advances in machine learning techniques that could help users navigate, explore, discover and understand interesting and useful content in books. Using a group of ten students and over one thousand crowdsourced judgments, we conducted multiple user studies to evaluate topics and related passages in books, all learned by topic modeling. Using ten books, selected from humanities (e.g. Plato's Republic), social sciences (e.g. Marx's Capital) and sciences (e.g. Einstein's Relativity), and four different evaluation experiments, we show that users agree that the learned topics are coherent and important to the book, and related to the automatically generated passages. We show how crowdsourced evaluations are useful, and can complement more focused evaluations using students who have studied the texts. This work provides a framework for (1) learning topics and related passages in books, and (2) evaluating those learned topics and passages, and moves one step toward automatic annotation to support topic navigation of books.

Categories and Subject Descriptors
H.4.0 [Information Systems]: General

Keywords:
topic modeling, book navigation, text mining

INTRODUCTION
In Plato’s Republic, what place does justice have in virtue? In Marx’s Capital, in what sense can we say that nature does not produce value; only labor produces value? These are typical questions asked of students studying these texts. To answer these questions a student may page through a copy of the text, looking for relevant passages. The student may also make use of the book’s index or table of contents.

With electronic books, students can conduct searches for key terms or phrases, and possibly find key passages to help them answer these types of essay questions. But the searching the Republic for ‘justice’ produces over 700 hits. A list of 700+ search results, or even a visualization of these hits within the book, barely scratches the surface of what could be done to provide effective navigation.

In this work we investigate whether topic modeling can automatically learn useful topics and related passages in books. Topic modeling is a technique for automatically extracting semantic themes from document collections, and has been shown to produce highly coherent and relevant topics. Here we focus on the value and utility of these learned topics. Since “useful” depends on the need or task, we use students studying the texts as a context for our evaluation.

METHODOLOGY
We present users with topics learned by the topic model, and related passages. In addition to exploring the degree to which users approved of topics produced by the topic model, we wanted to find out if there were important topics that the topic model missed. To provide implicit context for the identification of topics, we worked with students from the University of Michigan who had been enrolled in courses in which the texts used in the study were taught. To gain insight into the degree to which user characteristics inform approval ratings, we broadened the study to include crowdsourced workers.

Selection of Texts: Our criteria for selecting texts were that they be widely taught and that they be long enough to benefit from topic navigation. Additionally, for easier access, we required all texts to be out-of-copyright. We identified a list of ten authors/books in the sciences, social sciences, and humanities. Table 1 shows the ten authors, separated into the three used for the evaluation conducted with University of Michigan students, and then using the set of ten books for our study with crowdsourced workers.

<table>
<thead>
<tr>
<th>Author</th>
<th>Evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Book-of-Interest)</td>
<td></td>
</tr>
<tr>
<td>Plato (Republic)</td>
<td>6 students / crowdsourced</td>
</tr>
<tr>
<td>Marx (Capital – Part 1)</td>
<td>2 students / crowdsourced</td>
</tr>
<tr>
<td>(The Protestant Ethic and the Spirit of Capitalism)</td>
<td>2 students / crowdsourced</td>
</tr>
<tr>
<td>Plato, Marx, Weber, Darwin, Keynes, Mendel, Newton, Malthus, Einstein, Wollstonecraft</td>
<td>crowdsourced</td>
</tr>
</tbody>
</table>

Table 1. Books used in the student evaluation.
**Topic Modeling**

We used standard LDA topic modeling [1], learning 20 topics in every case. Each book was split into 1-page segments and modeled as part of a collection of similar texts by the same author or as a collection itself. We used the conventional approach of representing a topic with its top-10 terms. Table 2 shows one selected example topic from each of the ten books.

<table>
<thead>
<tr>
<th>Author</th>
<th>Example Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plato</td>
<td>just justice man good injustice unjust friend</td>
</tr>
<tr>
<td>Marx</td>
<td>money gold price circulation commodities silver</td>
</tr>
<tr>
<td>Weber</td>
<td>life asceticism puritan spirit rational ethical self</td>
</tr>
<tr>
<td>Darwin</td>
<td>species varieties distinct genus difference genera</td>
</tr>
<tr>
<td>Keynes</td>
<td>export percent goods import trade million raw</td>
</tr>
<tr>
<td>Mendel</td>
<td>white colour red dominant recessive type coloured</td>
</tr>
<tr>
<td>Newton</td>
<td>moon tide sea hour force after water height</td>
</tr>
<tr>
<td>Malthus</td>
<td>population country increase food number birth mortality age</td>
</tr>
<tr>
<td>Einstein</td>
<td>ordinate space continuum system four dimensional point event</td>
</tr>
<tr>
<td>Wollstonecraft</td>
<td>woman man behaviour kind men reputation respect</td>
</tr>
</tbody>
</table>

Table 2. Example topic from 20 topics learned from each of the ten books used in the crowdsourced evaluation.

In the above topics, the words are listed in order of importance to the topic. The topics learned by the topic model, while being relatively coherent, sometimes take some interpretation. For example, the topic from Mendel’s *Principles of Heredity* “white colour red dominant recessive type coloured” relates to the idea of Mendelian genetics and the transmission of characteristics (e.g. experiments crossing flowers of different colors). Likewise, the Einstein *Relativity* topic of “ordinate space continuum system four dimensional etc.” relates to Einstein’s four-dimensional formulation of space-time, one of several topics discussed in *Relativity*. Assuming some familiarity of these books, one could suggest that these topics have an instant face validity of being both relevant and important to the book. Our next step was to evaluate this suggestion.

**EXPERIMENTAL DESIGN**

The experiments used the ten books from Table 1, and two groups of evaluators (students and crowdsourced workers). The students had studied one book from Plato, Marx and Weber, and crowdsourced workers came from Amazon Mechanical Turk (AMT). The experiments were as follows:

- **Student-Eval-1**: What topics are in this book? (Students -- Plato/Marx/Weber)
- **Student-Eval-2**: Evaluate learned topics/passage for book (Students -- Plato/Marx/Weber)
- **Crowd-Eval-1**: Evaluate learned topics/passage for book (Crowd -- Plato/Marx/Weber)
- **Crowd-Eval-2**: Select which topic better relates to passage (Crowd -- Ten books)

**Student-Eval-1: What topics are in this book?**

Participants were assigned to one of three texts by Plato, Marx, and Weber that they had previously studied in a course. They were then asked to identify ten or more major topics, ideas, or themes discussed in the text.

**Student-Eval-2: Evaluate learned topics for book**

Participants were asked to complete a questionnaire on topics produced by the topic model for the text they had been assigned in Student-Eval-1. We took the fifteen or so topics that most related to the book of interest, and conducted the evaluation on those topics. For each of these topics, participants were presented with the top-10 words in the topic and a sample passage identified by the model as being about the topic. Students were then asked to rate their agreement with the following statements using a 5-point scale, ranging from *Strongly Agree* to *Strongly Disagree*.

1. The topic represents a coherent concept.
2. The topic is important in the book.
3. The topic accurately describes the passage.

**Crowd-Eval-1: Same as Student-Eval-2 run on AMT**

This task was identical to Student-Eval-2, with the following exceptions: (1) each item was presented individually to workers as a separate Human Intelligence Task (HIT); and (2) the second question on the importance of the topic in the book was omitted given that we could not assume AMT workers had read the book in question.

**Crowd-Eval-2: Which topic better relates to passage?**

The Crowd-Eval-2 task was designed to evaluate the topics and passages automatically learned for our set of ten books. While our long-term goal is to create and evaluate topic segmentation and topic navigation of books, this task addresses the smaller question of evaluating the coherence of topics, and the relevance of automatically learned topics to automatically selected matching passages. For the mini-collections, we needed to first identify topics that related to the book-of-interest. Here, with just the single book, we performed our evaluation over all twenty learned topics (and matching passages) for each of the ten books, with the entire evaluation generated in a fully-automated fashion. This comprehensive approach reduces potential bias in the evaluation, and gives a realistic evaluation of the range of quality one could expect from topic modeling.

**EXPERIMENTAL RESULTS**

**Student-Eval-1: What topics are in this book?**

Ten students from the University of Michigan participated in the student evaluations. Participants were selected on the basis of having previously studied one of the texts used in the study (Plato, Weber, Marx) and worked with the same text in Student-Eval-1 and Student-Eval-2.

In Student-Eval-1, participants were asked to identify major
topics discussed in the text and give a representative passage about each topic. Participants were asked to use a short phrase to describe each topic. Because the students had previously studied the texts in a course, we expected some agreement on important themes but also some variation in emphasis and expression. We used the sample passages to help identify overlapping topics.

We assembled responses from the six students studying Plato, and the learned topics. While there was broad agreement among participants on several general topics, there was also considerable variation in how students described their version of the topic. For example, two of the six students suggested there was a topic relating to *City*; however, one provided the more specific *Just City*. For this general topic, the topic model had three potentially matching topics, each emphasizing a distinct aspect of cities (law and justice, war and security, economy and trade). We see broad overlap in the learned topics and topics suggested by students that echo some of the major themes emphasized in the teaching of *The Republic* (the theory of forms, the immortality of the soul, the notion of philosopher-kings). It is also interesting to note that in some cases the topic model learned topics that were not identified by any of the students (e.g., Pleasure/Pain) and vice versa (e.g., Specialization).

We observed similar results for Marx and Weber, with the topic model learning approximately 80% of the topics suggested by the students (for all of Plato, Marx and Weber), based on there being semantic overlap in key terms or phrases.

The bar graphs in Figure 1 show mean student scores for the three questions relating to coherence, importance-to-book, and relevance-to-passage. Note that Marx and Weber only have N=2 students. The error-bars show standard error (when making comparisons, one could consider a difference significant if the respective error bars do not overlap). Our first observation is that the students on average agree that the learned topics and passages are coherent / important / relevant (the red dashed line at Score=3 indicates a neutral judgment). The results indicate that students found the topics for Weber slightly better than those for Plato, which in turn were better than the Marx topics.

Figure 2 gives a comparison of evaluations from students and crowdsourced workers. For Marx, there is no significant difference between students and workers in their evaluations. However, for both Plato and Weber, we see that the students scored both coherence and relevance higher than the workers. One possible explanation is that some of the words in a topic make sense only in the overall context of the passage in the book. Thus a worker who has not read the book might not infer as much meaning from just the passage, and therefore give lower ratings to the topic. While there are some difference in the results from students and workers, this could possibly be more of a calibration issue, particularly given that the students have studied the texts in a classroom setting.

Crowd-Eval-2: Evaluation of topic/passage for ten books
We crowdsourced judgments of 200 topics and passages (20 topics for each of the ten books). After omitting responses from spammers and other sub-standard workers who failed too many traps, we obtained a total of 1020 non-spam accepted responses, averaging five judgments per topic.

Our first result from Crowd-Eval-2 is shown in Figure 3. This Figure shows the accuracy or proportion of the time that workers selected the topic model topic in preference to the random list of words (given the choice between the two). We see this accuracy (of correctly identifying the topic) range from 64% for Einstein, up to 98% for Darwin, and approximately 80% averaged over all the ten books. This
result is quite promising: It says that 4 out of 5 times the automatically learned topics are identified as being more related to the passage than a set of ten words taken from the passage. Or -- said more strongly -- that workers do indeed think that the topics coherently relate to the passage.

But how difficult was this task? On average, the learned topic has anywhere from 6-8 out of the 10 words appearing in the passage (with an average of 6.7 matches). But the random topic, by design, always has all 10 words appearing in the passage, making this task more cognitively involved than simply counting word matches.

DISCUSSION AND CONCLUSION
We performed a series of experiments and evaluations of topics learned by a topic model of books, both using a group of ten knowledgeable students, and using more than one thousand judgments from crowdsourced workers. In conclusion, this work has shown that (1) topic modeling is able to learn sensible and usable topics and related passages in books, suitably granular for within-document navigation of longer texts; (2) user evaluation is important and should be incorporated early, possibly even before applications are developed, and (3) crowdsourced evaluations can enable inexpensive and rapid evaluation and prototyping.

Figure 3. Percent of time worker selected learned topic over random list of words (N=1020).

We also asked workers to score (on a 1 to 5 scale) how well their selected topic relates to the passage. On average, workers scored the learned topics 17% better than the random list of words, showing that workers think the learned topic better relates to the passage than the random list of words.

Looking forward: Illustration of how topics could be used for user navigation in books

While these experiments laid groundwork for topic navigation, we can illustrate how the topic modeling approach might be used in a topic navigation tool in an appropriate user interface (e.g. compare to [2]). Figure 4 shows topic trends from Plato’s Republic and Darwin’s Origin of Species. The figure shows the relative importance of selected topics throughout the course of the book. Not only do we see that the topic model provides page ranges that are relevant to a particular topic, using Origin of Species we see validation with existing chapter boundaries, in this case the “struggle” topic clearly matching up with Chapter 3: Struggle for Existence. One can imagine a user interface where the user has this type of graphical display and moves a slider to quickly flip to pages of interest, relating to some selected topic.

ACKNOWLEDGMENTS
This study was funded by a National Leadership Grant (LG-06-08-0057-08) from the Institute of Museum and Library Services.

REFERENCES