SYLLABI STUDIES TAKE TOO LONG!: AUTOMATING A TEXT DATA MINING APPROACH TO INSTRUCTIONAL OUTREACH

AMY JAMES, JOSHUA BEEN AND BETH FARWELL

INTRODUCTION

Curriculum mapping can help to identify your library’s instructional reach as well as reveal potential opportunities for increasing instruction and new information literacy partnerships with faculty and instructors across your campus. Sarah LeMire and Stephanie Graves (2019) provide the following explanation of curriculum mapping:

Curriculum mapping is a technique that visually represents curricular data, such as learning outcomes, level of instruction, and library instruction statistics, to improve program analysis. Libraries commonly use it as a method for illustrating connections between library instruction and the university curriculum, determining overlaps in library instruction, and identifying opportunities for increasing instruction. (p. 274)

Baylor University Libraries’ Director of Instruction and Information Literacy was interested in using curriculum mapping as a tool to work with individual liaison librarians from within the Research and Engagement department to determine opportunities for engagement with faculty in their departments through information literacy instruction. During the fall of 2020, the Director of Instruction and Information Literacy partnered with the Director of Data and Digital Scholarship and the Director of Arts and Special Collections Research Center to embark on a long-term curriculum mapping syllabi study. We decided to start with gathering syllabi from the Museum Studies department because of its size and because of the close ties the department has with the libraries.

GETTING STARTED

The insight that we hoped to gain from this study wasn’t much different from the prominent syllabi studies historically done in academic libraries (see works cited). However, our project differed because we planned to automate the process. The library liaison to the Department of Museum Studies worked with us to come up with some language that we would use with faculty to describe the project, emphasizing that we would not be evaluating or critiquing the syllabi or their instruction. We explained that we were looking for opportunities where we could provide support to faculty and instructors through information literacy instruction.

Then, we crafted a list of information literacy terms and concepts as well as indicators for levels of information literacy that each term was associated with in relationship to our library’s tiered approach to information literacy instruction. The terms that we selected came from scanning a variety of different course syllabi, looking at assignments, and scanning for important vocabulary within the ACRL Framework for Information Literacy.

The Director of Data and Digital Scholarship created a Python script to mine the syllabi for the pre-defined information literacy related terms and concepts. The algorithm works by identifying, based on their location, two components: 1) action verbs in context with 2) the research-related nouns within ten words from the verb(s). The second component is mapped to one of the three tiers of information literacy based on the verb found and it does not consider the noun. This technique allowed us to specify which level of information literacy the syllabi may have referred to. We selected the following nouns: library, collection, data,
project, paper, source, archive, theory, and component. We selected these verbs: evaluate, gather, synthesize, research, analyze, explore, identify, examine, and apply.

We created an interactive data visualization that liaison librarians can use as a tool to help provide targeted information literacy instruction outreach within their departments. The data visualization tool not only links to the syllabi collected but is also connected to the library’s instruction statistics. This allows the liaison to filter the data to see which classes in their department have information literacy concepts within their syllabi, which of those classes have received information literacy instruction in the past two years, and which ones do not have information literacy concepts in them at all (and therefore may not be good candidates for information literacy instruction).

COLLECTING SYLLABI

One of the first challenges we encountered in this project was gathering syllabi. Baylor University does not maintain a repository of syllabi and there is no university-wide expectation to collect them. To test the tool, we needed a handful of syllabi to start. We approached the Museum Studies Department as a pilot group based on the relatively small number of faculty and strong ties to the library through several projects. We received eleven syllabi to start. The School of Education was our next area to approach. We were not able to collect very many from this group, despite multiple attempts through emails to key faculty, notification sent through the department head, and numerous follow ups.

Our next approach was to investigate mining the syllabi housed in Canvas, our course management system. After several conversations with IT and others, we received initial permission to use this data—but this plan ultimately was not approved with the rationale that we did not have permission from faculty to use their content. We do have hope that if our initial project results are successful, we will be able to approach this option again.

We were discouraged that we could not work in Canvas, but soon we received a big break through the Arts & Sciences Core Curriculum group who have gathered syllabi for all courses in the core. The Arts and Sciences’ Dean’s office signed off on using these syllabi for our project with the caveat that we do not publish any information that might identify a particular faculty member.

After this good news, we also learned that one of our liaisons discovered the Department of Religion keeps current syllabi and we received approval to use two years of content.

RESULTS

While we have access to library instruction data with the Museum Studies department, there was still a lot about their courses and assignments that we did not know. Using Microsoft Power BI allowed us to see combined data points and understand where activity was occurring with possible information literacy goals. After a review of the results, we decided to focus on two classes.

The first course we looked at was MST 5318 (Ethical Issues in Museum Collections Management). We wanted to work with a course that would be well-suited to information literacy support, but as a graduate course had room for additional information literacy support. This course is a graduate course that was previously combined with an undergraduate section. The course will now be taught just as a graduate course and this is an ideal time to build in information literacy research modules as the professor is building course objectives. We are currently meeting with the professor to talk through options.

The second course we decided to work with was MST 5311 (Issues in Museum Administration). This course did not show up at all in the results—no instruction, and no matching terms. We were tempted to ignore this course, but knew it was one of the core courses required for the degree, so we decided to take a closer look. There is a required paper, and our plan is to build a library basics module.

While there is still work to be done in identifying the best path forward with departments, this tool gives liaisons a portal into a department’s activity. We were able to physically see gaps and opportunities for more partnerships in building information literacy instruction.

LOGISTICS

Automatically identifying new instruction opportunities for subject liaisons is a two-step process. First, the syllabi must be mined for IL concepts. To accomplish this, we created a publicly available and open-source Python Jupyter Notebook. This Notebook, named Syllabi Information Literacy Miner (SILM), is hosted on Google Colab and can be used directly through a web
browser with no installation or coding required. Second, the output from SILM is combined with our Springshare LibInsights instruction database in an interactive Power BI data report. This data report allows our subject liaisons to identify syllabi with IL components that we have not done instruction for. However, as it is configured for the specific format of Baylor University’s LibInsights database, this tool is not available to the public.

**Syllabi Information Literacy Miner (SILM)**

SILM is a publicly available and open source Jupyter Notebook. The digital object identifier (DOI) url is https://doi.org/10.18738/T8/EYYX7L. A Jupyter Notebook is an interactive Python environment where code, data, and multimedia can be analyzed, visualized, and documented in a web browser. Google Colab provides free Jupyter Notebook functionality. As the Colab in the name implies, it supports collaboration and sharing. While we coded SILM in Python, this tool can be used online without touching or even seeing any of the code. To see the underlying Python code, simply click the more options button on the top right corner and click Form/Hide Form.

The primary Python libraries used in SILM are textract (https://textract.readthedocs.io/en/stable/) and antiword (http://www.winfield.demon.nl/) to convert provided syllabi to text, pandas (https://pandas.pydata.org/) to work with structured tables, and matplotlib (https://matplotlib.org/) to generate visualizations.

SILM goes through five steps to identify IL concepts in syllabi. First, syllabi can be provided by uploading or providing a URL to a .zip archive. Syllabi can be provided in PDF, Word (.docx or .doc), or text file formats. The Notebook will then convert all files to text. After converting files to text, IL components are identified by locating specified verbs within context or within proximity to specified nouns. Next, the IL tiers of Library Basics, Library Research, and Research in the Disciplines are assigned based on the identified verb. At this point, visualizations are generated, and a downloadable Excel report is automatically downloaded.

There are five configurable settings on the Jupyter Notebook, just beneath the documentation. We set these default settings as optimized for mining Baylor University syllabi, but just about everything can be adjusted. The first time someone uses SILM, a copy of the Notebook is automatically saved to the user’s Google Drive, so any adjustments to the settings will apply solely to the user’s copy. The six configurable settings are (1) upload syllabi or provide link to url, (2) action verbs, (3) nouns that must be within context of the action verbs, (4) which verbs align with each IL tier, and (5) context size.

When completed, four visuals are available directly below the output, along with an automatically downloaded Excel file. The first visual is a pie chart showing the proportion of IL components found in each tier. Next, a column chart is created showing the counts of verbs located with a noun within context. Next is a displayed table showing the uploaded syllabi and which tiers were identified in each syllabi. Next, the most granular output is displayed. This is a table with each row representing an identified IL component. The first column is the identified verb. Next the file. Next the full context. The last three columns specify which tier the verb is aligned with. Finally, this granular table is downloaded as an Excel file.

**Power BI Data Report (Baylor University Only)**

The Microsoft Power BI-generated data report is the web application that subject liaisons at Baylor University will use to identify new opportunities. Its purpose is to connect the IL components identified using SILM with recorded instruction events as stored in Baylor University’s Springshare LibInsights database so that courses, that the library has not recently provided instruction, can be identified with IL outcomes. At the time of this LOEX presentation, this data report contains two years of Core Curriculum syllabi from the College of Arts & Sciences, Department of Museum Studies syllabi, School of Education syllabi, and Department of Religion syllabi.

Power BI is a business intelligence tool used to visualize analytics and dashboards. The data report is available via a web browser. Power BI is directly connected to Baylor’s Active Directory, so liaison librarians must authenticate in order to view and interact with the report.

The data report uses a matrix visual to provide subject liaisons with the ability to drill down and explore for new instruction opportunities. The department and course number of syllabi are listed with identified Information Literacy components. A count of the number of IL components found per course for each tier are listed in the next three columns. In the final column, the date of the most recent library instruction session for that course is listed. The listed course can be drilled down to the identified verbs, the full context, and the full syllabi file name. Additionally, subject liaisons can export the content to a downloadable Excel file. This Excel file is similar to the output from the Jupyter Notebook, but also includes the date of the last instruction.
Power Query M, the preprocessing language available in Power BI, is used to clean and combine the output from the Jupyter Notebook and the LibInsights instruction data. Power Query M is analogous to OpenRefine for Power BI data models. This query is automatically repeated with any data update or refresh.

LESSONS LEARNED

Once we began trying to collect syllabi from individual departments, we realized the response rate would not supply data at a scale we had anticipated. This prompted us to begin looking at additional ways to harvest syllabi and resulted in acquiring access to the Arts and Sciences Core syllabi.

Because Baylor is a private institution, syllabi are not collected or made available internally or externally. This was the biggest challenge that we encountered. Many faculty that we spoke to about the project expressed concerns about where the syllabi might be shared, despite us reassuring them that they would not be used for anything outside the project scope. We were able to work with the Arts and Sciences Core to gain access to the syllabi that were already collected. That involved an approval process, and it took the library Dean writing the Arts and Sciences Core Director to ensure that the syllabi wouldn't be used for anything else. We also have discovered through our Religion & Theology liaison that the religion department maintains a Box folder with syllabi over several semesters. We were able to speak directly with the department chair and gain access to those folders as well through the established liaison relationship.

REFERENCES


APPENDIX

Below: Sample Output from SILM using 112 Museum Studies Syllabi

Analyzing Content
>> Downloading and Unzipping Syllabi
>> Converting Syllabi to Text
>> Identify Terms
>> Write to Excel

Proportion of Information Literacy Levels Across Syllabi

Information Literacy Categories Identified in Syllabi

- Library Basics: 31.6%
- Research Basics: 34.2%
- Research in the Disciplines: 40.3%

Information Literacy Verbs Identified Across Syllabi

Terms (Verbs) Found in Syllabi

- research: 20
- gather: 10
- evaluate: 10
- identify: 10
- explore: 10
- apply: 5
- analyze: 5
- examine: 5
- synthesize: 5

Below: Power BI Data Report Screenshot
### Syllabi Information Literacy Miner with Date of Last Instruction

<table>
<thead>
<tr>
<th>Dept and Department</th>
<th>Library Basics</th>
<th>Research Basics</th>
<th>Research in the Disciplines</th>
<th>Most Recent Library Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1301</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2310</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4301</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>CLA (Classics)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1304</td>
<td>0</td>
<td>15</td>
<td>20</td>
<td>01/23/20</td>
</tr>
<tr>
<td>2306</td>
<td>0</td>
<td>4</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>3301</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>CSS (Communication)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1302</td>
<td>3</td>
<td>10</td>
<td>20</td>
<td>08/31/20</td>
</tr>
<tr>
<td>ECO (Economics)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1305</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>10/19/20</td>
</tr>
<tr>
<td>EDC (Curriculum and Instructions)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6345</td>
<td>0</td>
<td>7</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>